Bayesian Multi-Model Frameworks

Properly Addressing Conceptual Uncertainty in Applied Modelling

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Marvin Höge
aus Mühlacker

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	25.2.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr.-Ing. Olaf Cirpka
2. Berichterstatter:	Prof. Dr.-Ing. Wolfgang Nowak
3. Berichterstatter:	Dr. Thomas Wöhling

Abstract

We use models to understand or predict a system. Often, there are multiple plausible but competing model concepts. Hence, modelling is associated with conceptual uncertainty, i.e., the question about proper handling of such model alternatives. For mathematical models, it is possible to quantify their plausibility based on data and rate them accordingly. Bayesian probability calculus offers several formal multi-model frameworks to rate models in a finite set and to quantify their conceptual uncertainty as model weights. These frameworks are Bayesian model selection and averaging (BMS/BMA), Pseudo-BMS/BMA and Bayesian Stacking.

The goal of this dissertation is to facilitate proper utilization of these Bayesian multi-model frameworks. They follow different principles in model rating, which is why derived model weights have to be interpreted differently, too. These principles always concern the model setting, i.e., how the models in the set relate to one another and the true model of the system that generated observed data. This relation is formalized in model scores that are used for model weighting within each framework. The scores resemble framework-specific compromises between the ability of a model to fit the data and the therefore required model complexity.

Hence, first, the scores are investigated systematically regarding their respective take on model complexity and are allocated in a developed classification scheme. This shows that BMS/BMA always pursues to identify the true model in the set, that Pseudo-BMS/BMA searches the model with largest predictive power despite none of the models being the true one, and that, on that condition, Bayesian Stacking seeks reliability in prediction by combining predictive distributions of multiple models.

An application example with numerical models illustrates these behaviours and demonstrates which misinterpretations of model weights impend, if a certain framework is applied despite being unsuitable for the underlying model setting. Regarding applied modelling, first, a new setting is proposed that allows to identify a "quasi-true" model in a set. Second, Bayesian Bootstrapping is employed to take into account that rating of predictive capability is based on only limited data.

To ensure that the Bayesian multi-model frameworks are employed properly and goal-oriented, a guideline is set up. With respect to a clearly defined modelling goal and the allocation of available models to the respective setting, it leads to the suitable multi-model framework. Aside of the three investigated frameworks, this guideline further contains an additional one that allows to identify a (quasi-)true model if it is composed of a linear combination of the model alternatives in the set.

The gained insights enable a broad range of users in science practice to properly employ Bayesian multi-model frameworks in order to quantify and handle conceptual uncertainty. Thus, maximum reliability in system understanding and prediction with multiple models can be achieved. Further, the insights pave the way for systematic model development and improvement.

Kurzfassung

Wir benutzen Modelle, um ein System zu verstehen oder vorherzusagen. Oft gibt es dabei mehrere plausible aber konkurrierende Modellkonzepte. Daher geht Modellierung einher mit konzeptioneller Unsicherheit, also der Frage nach dem angemessenen Umgang mit solchen Modellalternativen. Bei mathematischen Modellen ist es möglich, die Plausibilität jedes Modells anhand von Daten des Systems zu quantifizieren und Modelle entsprechend zu bewerten. Bayes'sche Wahrscheinlichkeitsrechnung bietet dazu verschiedene formale Multi-Modellrahmen, um Modellalternativen in einem endlichen Set zu bewerten und ihre konzeptionelle Unsicherheit als Modellgewichte zu beziffern. Diese Rahmen sind Bayes'sche Modellwahl und -mittelung (BMS/BMA), Pseudo-BMS/BMA und Bayes'sche Modellstapelung.

Das Ziel dieser Dissertation ist es, den adäquaten Umgang mit diesen Bayes'schen Multi-Modellrahmen zu ermöglichen. Sie folgen unterschiedlichen Prinzipien in der Modellbewertung weshalb die abgeleiteten Modellgewichte auch unterschiedlich zu interpretieren sind. Diese Prinzipien beziehen sich immer auf das Modellsetting, also darauf, wie sich die Modelle im Set zueinander und auf das wahre Modell des Systems beziehen, welches bereits gemessene Daten erzeugt hat. Dieser Bezug ist in Kenngrößen formalisiert, die innerhalb jedes Rahmens der Modellgewichtung dienen. Die Kenngrößen stellen rahmenspezifische Kompromisse dar, zwischen der Fähigkeit eines Modells die Daten zu treffen und der dazu benötigten Modellkomplexität.

Daher werden die Kenngrößen zunächst systematisch auf ihre jeweilige Bewertung von Modellkomplexität untersucht und in einem entsprechend entwickelten Klassifikationschema zugeordnet. Dabei zeigt sich, dass BMS/BMA stets verfolgt das wahre Modell im Set zu identifizieren, dass Pseudo-BMS/BMA das Modell mit der höchsten Vorsagekraft sucht, obwohl kein wahres Modell verfügbar ist, und dass Bayes'sche Modellstapelung unter dieser Bedingung Verlässlichkeit von Vorhersagen anstrebt, indem die Vorhersageverteilungen mehrerer Modelle kombiniert werden.

Ein Anwendungsbeispiel mit numerischen Modellen verdeutlicht diese Verhaltenweisen und zeigt auf, welche Fehlinterpretationen der Modellgewichte drohen, wenn ein bestimmter Rahmen angewandt wird, obwohl er nicht zum zugrundeliegenden Modellsetting passt. Mit Bezug auf anwendungsorientierte Modellierung wird dabei erstens ein neues Setting vorgestellt, das es ermöglicht, ein "quasi-wahres" Modell in einem Set zu identifizieren. Zweitens wird Bayes'sches Bootstrapping eingesetzt um bei der Bewertung der Vorhersagegüte zu berücksichtigen, dass diese auf Basis weniger Daten erfolgt.

Um zu gewährleisten, dass die Bayes'schen Multi-Modellrahmen angemessen und zielführend eingesetzt werden, wird schließlich ein Leitfaden erstellt. Anhand eines klar definierten Modellierungszieles und der Einordnung der gegebenen Modelle in das entspechende Setting leitet dieser zum geeigneten Multi-Modellrahmen. Neben den drei untersuchten Rahmen enthält dieser Leitfaden zudem einen weiteren, der es ermöglicht ein (quasi-)wahres Modell zu identifizieren, wenn dieses aus einer Linearkombination der Modellalternativen im Set besteht.

Die gewonnenen Erkenntnisse ermöglichen es einer breiten Anwenderschaft in Wissenschaft und Praxis, Bayes'sche Multi-Modellrahmen zur Quantifizierung und Handhabung konzeptioneller Unsicherheit adäquat einzusetzen. Dadurch lässt sich maximale Verlässlichkeit in Systemverständis und -vorhersage durch mehrere Modelle erreichen. Die Erkenntnisse ebnen darüber hinaus den Weg für systematische Modellentwicklung und -verbesserung.

Contents

List of Figures

List of Tables

List of Symbols

$p(\cdot)$ and $q(\cdot)$	Marginal (prior) probability density function of a random variable
$p(\cdot\|\cdot)$	Conditional (posterior) probability density function of a random variable
E[·] or E[·|·]	Expectation of a random variable (marginal or conditional)
Var[·] or Var[·|·]	Variance of a random variable (marginal or conditional)
$\boldsymbol{D}$	Within-sample/calibration data vector
$\boldsymbol{D}'$	Out-of-sample/validation data vector
$\boldsymbol{\epsilon}$	Measurement error vector
$\boldsymbol{\Theta}$	Model parameter vector from parameter space Ω_m
$\boldsymbol{x}$	Model input vector
$\boldsymbol{y}$	Model prediction/forecast/output vector from prediction space $\mathcal{Y}$
$\hat{\boldsymbol{y}}$	Best estimate (e.g., maximum likelihood) model prediction vector
M_m	Individual model m from model space $\mathcal{M}$
$\boldsymbol{M}$	Model set / ensemble
w_m	Model weight of model m
$\boldsymbol{w}$	Model weight vector
N_s	Number of observations
N_p	Number of parameters
N_p^*	Number of effective parameters
N_M	Number of models
$\mathcal{S}$	Model rating score
$\mathcal{F}$	Goodness-of-fit term
$\mathcal{C}$	Model Complexity term
$\mathcal{K}$	Convex set of models
$\ln(\cdot)$	natural logarithm
$\mathcal{N}(\mu, \sigma^2)$	Normal/Gaussian distribution with mean μ and standard deviation σ

List of Acronyms

BB	Bayesian Bootstrap
BCMS/BCMA	Bayesian combined model selection / averaging
BCV	Bayesian cross-validation
BME	Bayesian model evidence
BMS/BMA	Bayesian model selection / averaging
DGP	Data-generating process
D_{KL}	Kullback-Leibler divergence
DoF	Degrees of freedom
elpd	Expected logarithmic predictive density
EPE	Expected prediction error
GC	Geometric complexity
i.i.d.	Independent and identically distributed
LOO (CV)	Leave-one-out (cross-validation)
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MDL	Minimum Description Length
OF	Occam factor
PDE	Partial differential equation
pdf	Probability density function
Pseudo-BMS/BMA	Pseudo-Bayesian model selection / averaging
QoI	Quantity of interest
RMSE	root-mean-square error
WSSE	weighted sum of squared errors
xIC	x information criterion (e.g. AIC: Akaike information criterion)

1 Introduction

1.1 Motivation

Models and Uncertainty in Modelling

Whenever we want an explanation of the past, confirmation in the present or predictions of the future, we employ models. In most simple and general terms, a model is "a thing used as an example to follow or imitate" (Oxford), i.e., something that allows us to understand or forecast behaviour of the system we are interested in, whether its underlying relations are of physical, ecological, technical, political, social, economical, financial, or other nature. More specifically, for quantitative modelling of systems, models are rather to interpret as "a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions" (Oxford). Mathematical models enable us to simulate systems under all kinds of conditions, to perform scenario analyses and, ultimately, to support decision-making (Reichert et al., 2015; Ferré, 2017).

However, modelling is subject to uncertainty. The shear attempt of creating a model implies uncertainty due to simplifications, assumptions, etc. Operating a model adds and propagates uncertainty further due to all kinds of errors, vague included observations, etc. Therefore, modelling comes with an intrinsic demand for uncertainty quantification. Numerous approaches were proposed about how to quantify uncertainty in model results based of sensitivity analyses (Gupta and Razavi, 2018), interval computation (Moore, 1979), fuzzy set theory (Zadeh, 1965; Klir and Yuan, 1995) and possibility theory (Zadeh, 1999) or entropy (Shannon, 1948), naming just a few. Yet, the most thoroughly discussed and widely used tool to quantify uncertainty is probability theory (Gillies, 2012).

In particular, Bayesian inference proved to be suitable to handle uncertainty because under the Bayesian paradigm probability resembles belief that quantifies lack of knowledge (e.g. Bernardo and Smith, 1994; Gelman and Shalizi, 2013), i.e., the paramount origin of uncertainty in modelling. Under this interpretation, Bayesian probability theory remains as the only consistent mathematical calculus for uncertainty quantification (Nearing et al., 2016, and references therein). Plainly, mathematical models contain parameters that map a certain input to a corresponding output. Bayesian inference provides a comprehensive framework that allows to address the vague notion of overall model uncertainty, e.g., by decomposing it accordingly into parameter, input, output and so-called conceptual uncertainty (Renard et al., 2010; Schöniger et al., 2014).

Multiple Models and Conceptual Uncertainty

While parameter, input and output uncertainty are evaluated for individual models, they can be subsumed under the overarching conceptual uncertainty, i.e., the question of how to deal with multiple alternative models (hypotheses). This uncertainty goes beyond *how the parameters can map input to output more reliably* to *whether the chosen parametrization is adequate at all.* It defines which inputs, parameters and outputs are relevant and targets at the very core of science: systematic hypothesis testing and inductive inference, i.e., how to infer general rules from special cases and observations (Rathmanner and Hutter, 2011).

This issue about model choice dates back to ancient philosophy where Epicurus stated the principle of multiple explanations and, concomitant, to never discard a plausible hypothesis when it is consistent with the observations (e.g. Rathmanner and Hutter, 2011). The multi-model approach stands to reason also in light of the Duhem-Quine thesis (see Harding, 1975), i.e., that any single model suffers from underdetermination by observations. Thereafter, a single hypothesis cannot be isolated and evaluated because it always relies on other (auxiliary) hypotheses or assumptions. Hence, multiple models that utilize divergent auxiliaries illuminate a system from different angles and thereby shed light on shortcomings of individual models.

Employing multiple models requires rules to rate their ability to "follow or imitate" to the modelers desire. Some models are more plausible than others and objective model rating prevents us from unreasonable preference of one model over others (Elliott and Brook, 2007). However, a rationale for such rules and formal rating is not straight-forward. A qualitative heuristic is given by Occam's razor (e.g. Hutter, 2007, and references therein). It states that between competing hypotheses, the model that needs fewest assumptions (is the simplest model) while still being consistent with the observations is best. A model that follows this principle is called parsimonious (Angluin and Smith, 1983).

In our era of machine-based computation, Solomonoff (1964) formalized these principles for the first time entirely by a concept called algorithmic probability (Rathmanner and Hutter, 2011). The basic assumption is that our observations were generated by some "true" algorithm. In theory, it is then possible to systematically rate (Occam's razor) all imaginable hypotheses (principle of multiple explanations) by writing them down as algorithms and computing them on a Universal Turing Machine (Rathmanner and Hutter, 2011), i.e., the most general platform for algorithm execution (Grünwald and Vitányi, 2003). Solomonoff's algorithmic probability then allows to translate the code-length of each algorithm into a Bay-

esian probability. This way, it offers (theoretically) a way to place a probability distribution over the entirety of potential models (Hutter, 2007). Each algorithm resembles a compression of the observations - storing each observation separately resembles the longest possible code. Hence, the probability of an algorithm to be the optimal compression is the higher the shorter its code-length is. The best model, in this spirit, is the shortest code that can fully reproduce the observations and then halts (Grünwald and Vitányi, 2003) - and this is the true data-generating algorithm. This comprehensive approach is called universal induction. It is theoretically complete, but practically incomputable (Rathmanner and Hutter, 2011).

Despite its practical inapplicability, the above concept sheds light on all issues of conceptual uncertainty in practice. Three principal issues and their consequences can be elicited:

1. The range of possible models is huge or even infinite - in theory, we need to check all alternatives to know which one is absolutely the best. In practice, we deal with a subset which is known as *finite hypotheses problem* (Nearing et al., 2016). Therefore, it is by no means guaranteed, that one of our models is the algorithm that generated the data at all. Hence, we need a system to qualitatively describe the relation of this finite subset toward the so-called data generating process (DGP).

2. In scenarios where a *true model of the DGP is unavailable*, aiming at identifying it is no suitable objective. Then, conceptual uncertainty does not relate to the probability of being the true model any more. As a consequence thereof, interpretations of conceptual uncertainty within such scenarios have to be adjusted. Subsequently, Occam's razor for model rating and selection also needs to be realized by different means than algorithmic probability because the underlying assumptions are different, too.

3. Under the assumption that a true model cannot be found, it is questionable whether model rating with the ultimate goal to select only one model is promising. Rather, we require systematic approaches to *successfully operate multiple non-true models*.

In practice, we frequently consider multiple model alternatives (Clyde and Iversen, 2013) and therefore need guidance for appropriately matching "truth scenarios" and "razor implementations" in order to adequately rate and utilize a single member or multiple alternatives within a model set. Such guidance, however, is too rare and too little systematic in the available literature.

Model Types and Model Settings

For every modelling task at hand, it is typically possible to come up with several hypotheses that are consistent with the observations - theoretically, it is even "possible to propose an infinite number of different models that allow us to correctly predict any finite number of events" (Nearing et al., 2016) if we had infinite time. Looking only at some model classification approaches reveals that numerous types of models and kinds of modelling schools exist: physics-based vs. data-driven, linear vs. non-linear, deterministic vs. stochastic, slow vs. fast, etc. (see, e.g., Refsgaard and Abbott, 1996; Breiman, 2001). Models fall into several of such categories, e.g., physics-based and stochastic, and even within such a specification there are countless possible alternatives. Using expert knowledge, we are able to restrict this huge variety but still, we can often only guess whether we "get close" to the DGP with a certain model or not.

Settings that describe whether (a) the true model is within our set of candidate models, (b) the true model exists but is outside of our set or (c) the true model is per se unavailable have been proposed by Bernardo and Smith (1994) from a decision theoretic perspective. They enable us to formalize this issue and serve as starting point for any kind of successive model rating. Despite the difficulty of transferring such settings to practical modelling tasks, such a distinction is crucial before any method for model rating can be applied: Model rating methods that are tailored to identify a true model will yield misleading results if applied in a setting where the true model is not included. Likewise, model rating methods that assume the truth to be unavailable among the candidate models can also never support the claim that the model they rate best is probably the true model. As example, when models are used to gain understanding of an isolated process in natural sciences they need to represent the relevant physics at the process-scale and have to be rated by methods that can identify a true model.

Mostly, such settings are recognized and discussed in the field of statistics (e.g. Clyde and Iversen, 2013). This motivates to elaborate on their relation to specific model types in order to make them accessible to a broader audience for practical application in other fields. Existing literature on model rating in most scientific disciplines does not take these settings into account.

Model Complexity and Model Rating

In practice, we typically deal with a finite set of distinct and fully defined models that share at least one objective like a quantity of interest (QoI) and account for conceptual uncertainty between all competitors in this respect. The law of par-

simony manifests itself as trade-off between quantified fit of model predictions to observations and the vague notion of model complexity: A model is best if it fits observations at least equally well or better while being simpler than the alternative models. This implies that, for a given amount of data, there is an optimal complexity of a model (Claeskens, 2016), which is neither too complex, nor too simple (see Occam's razor).

Numerous model rating methods were developed that all employ this trade-off but deviate vastly in their rating results due to the vagueness about what model complexity actually is. Universal induction (Solomonoff, 1964) provided the first rigorous definition of model complexity. Model rating methods that are directly derived from it link model complexity to code length (Grünwald and Vitányi, 2003). Others measure model complexity by the number of (effective) parameters (Akaike, 1973; Spiegelhalter et al., 2002), the probabilistic distribution of parameters (Kashyap, 1982), the sensitivity of parameters to observations (Mallows, 1973), the spread of predictions (Friedman et al., 2001), mixtures of the previous or other means. There is no unique definition of model complexity and, often enough, model rating is only based on quantified fit obtained via an error metric like simple root-mean-square-error or so-called loss-functions without any consideration of model complexity. This is however insufficient regarding so-called model generalizability (e.g., Friedman et al., 2001), i.e., it does not allow to estimate the model performance for new data.

Among all those model rating methods, there are some that share the same underlying principles including similar representations of model complexity. Their asymptotic equality is often shown in the limit of infinite observations. However, in practice, there is only a limited amount of data. This limitation prevents that any hypothesis could ever be proven right (Popper, 2005; Tarantola, 2006) and exacerbates to assess which ones are good or bad (Nearing et al., 2016). Hence, the model rating methods require classification and interpretation with respect to the finite amount of data they are operated with and how they decide which models are better or worse than others. Attempts in this direction were made before (e.g. Kadane and Lazar, 2004; Yang, 2005; Vrieze, 2012), but are spread over many different scientific disciplines and typically compare distinct methods rather than aiming for an encompassing overview. A general classification scheme that, first, clearly depicts what is meant by model complexity, and, second, from which recommendations for action can be deduced for a specific modelling task at hand, is still missing.

Multi-model frameworks and their adequate utilization

Several multi-model frameworks are related to these model rating methods and allow for statistical model selection and averaging (e.g. Burnham and Anderson, 2004; Gelman et al., 2004). The most prominent example might be Bayesian Model Selection or Averaging (BMS or BMA; Draper, 1995; Hoeting et al., 1999; Raftery et al., 2005) in which model probabilities are used to express uncertainty between models in terms of how likely it is that a certain candidate model generated the observed data. Both BMS and BMA enjoy wide-spread usage over many disciplines (e.g. Trotta, 2008; Faust et al., 2013; Hooten and Hobbs, 2015; Schöniger, 2016) where they are often first choice to deal with conceptual uncertainty. Similarly, so-called Pseudo-BMA (Geisser and Eddy, 1979; Yao et al., 2017) is used to handle uncertainty between multiple models with respect to their individual ability to predict potential future data. In a similar spirit, model rating methods like the famous Akaike information criterion (AIC; Akaike, 1973, 1974) serve as basis for model selection. Other frameworks like (Bayesian) Stacking (Yao et al., 2017; Le et al., 2017) allow to combine model competitors in a set for predictive purposes rather than quantifying the uncertainty about one being relatively best.

Despite the popularity of multi-model frameworks, their applications frequently cause confusion: different model selection criteria rate completely different models as "best" (Burnham and Anderson, 2002; Claeskens et al., 2008); despite a true model possibly being in the set, criteria like the AIC are not able to identify it (Vrieze, 2012); model averaging by BMA often yields worse predictive results than single model use (Domingos, 2000; Clarke, 2003), which raises the question whether model combination as weighted average can be provided by BMA at all (see Minka, 2002) and how actual model combination can successfully be performed (Clyde and Iversen, 2013; Le et al., 2017).

All of these problems can be traced back to the **central question**: Which multi-model framework employs the adequate Occam's razor with respect to the model setting of the modelling task at hand? Then, among other insights, it becomes apparent that "BMA is no panacea" (Clyde and Iversen, 2013) and that conceptual uncertainty has different meanings and needs to be accounted for differently as well. Still, much too often the fundamental principles are neglected and multi-model methods are applied to practical modelling tasks decoupled from them. A thorough investigation on these principles that collects insights from various scientific disciplines and elicits guidance by highlighting linkages to method application is still missing.

Water and Hydrosystem Modelling

While the philosophical and statistical questions of conceptual uncertainty concern all scientific disciplines on a rather abstract level, their answers pertain to very practical consequences when applied in fields with direct impact on our every-day lives. In hydrosystem modelling, our interest is water and its ubiquitous impact: life rests on the availability of water; we humans depend on it for drinking and agricultural irrigation, hygiene and health care, or energy and industrial production. Water is vital to us and our cultural progress and requires sustainable management of water quality and quantity. At the same time, the excessive abundance of water during floods or harmful scarcity of water throughout droughts in the course of weather and climate-related events require prognoses and adaption. Hydrosystem models help us to deal with these issues on distribution and protection of water resources and the risk assessment of water-related threats.

Traditionally, hydro(geo)logists employ process-based models (Freeze and Harlan, 1969; Montanari and Koutsoyiannis, 2012) to gain system understanding and as primary choice to support decision making (Reichert et al., 2015) in addressing the above challenges. Thereby, a major concern is uncertainty that arises from simplification of the underlying physical processes (e.g. Renard et al., 2010; Clark et al., 2011; Refsgaard et al., 2012; Elshall and Tsai, 2014), which makes model complexity also a question of scale like temporal resolution or spatial variability and heterogeneity (Mendoza et al., 2015; Orth et al., 2015). As a consequence thereof the space of potential models expands by, e.g., lumped bucket-type (Bergström and Singh, 1995), spatially distributed (Refsgaard and Abbott, 1996), meso-scale hydrologic (Samaniego et al., 2010) models or even neural networks (Hsu et al., 1995) as modelling approaches on a completely different conceptual basis.

For decades, the hydro(geo)logic community has actively debated whether one approach is more suitable than others (e.g. Freeze and Harlan, 1969; Bergström and Singh, 1995; Blöschl and Sivapalan, 1995; Wagener et al., 2009; Mendoza et al., 2015). However, beyond individual preferences, there is no clear consensus on preferring one modelling approach over another when all approaches appear plausible for a certain task - and it remains questionable whether such a principle preference is justifiable. Nonetheless, there is consensus about the necessity to rigorously evaluate and rate models on a quantitative basis in order to justify an objective preference (e.g. Clark et al., 2008; Schöniger et al., 2014). Correspondingly, the growing insight of the community that "stochastification" of models allows for rigorous estimation of uncertainty for all kinds of hydrosystem models (Liu and Gupta, 2007; de Barros and Nowak, 2010; Cirpka et al.; Montanari and Koutsoyiannis, 2012; Nearing et al., 2016) spurred various attempts of so-called

Bayesian total error analyses (Kavetski et al., 2006; Vrugt et al., 2008; Reichert and Mieleitner, 2009; Renard et al., 2010; Montanari and Koutsoyiannis, 2012; Giudice et al., 2013) up to the level of conceptual uncertainty estimation using, e.g., BMA (Ye et al., 2010; Wöhling et al., 2015; Schöniger, 2016) or AIC-related model rating (Schoups et al., 2008). Thereby, however, the same shortcomings of model selection and averaging methods as mentioned above have been recognized (e.g. Poeter and Anderson, 2005; Ye et al., 2008; Lu et al., 2011; Schöniger, 2016).

1.2 This Thesis

1.2.1 Goals and Objectives

The goal of this thesis is to enable modellers to properly address conceptual uncertainty: For modelling tasks where multiple competing models are available, I explain and discuss which multi-model framework is most suitable to achieve a certain modelling goal that complies with the underlying model setting. Therefore, I deeply analyse theoretical underpinnings of these frameworks; I demonstrate their proper usage in applied modelling; I unify scattered insights from across various scientific disciplines that work on related topics; and I offer a map for the confusing field of conceptual uncertainty. Ultimately, I aim at making these multi-model frameworks more accessible, in particular to applied modellers that often face conceptual uncertainty in practice but are not (yet) familiar with the background required to address it.

Bayesian multi-model frameworks allow for proper uncertainty quantification in stochastic modelling. Within this thesis, I clarify how these Bayesian tools work and how they can be used successfully when modellers desire process-identification, predictive reliability and decision support in multi-model use.

This thesis is an attempt to close the highlighted gaps in Section 1.1 by bridging the theoretical and philosophical underpinnings of multi-model usage to applied modelling tasks. Therefore, I focus on examples from hydrosystem modelling but, without any loss of generality, the insights can be transferred to other disciplines where models are used, e.g., machine learning, psychology, ecology, engineering or economics.

1.2.2 Research Questions and Contributions

An effective utilization of model rating methods and built-on multi-model frameworks is complicated by the sprawling amount of alternatives and lack of guidance. Modellers are often forced to pick one method in order to put model selection or

averaging on a quantifiable basis - often this is a commonly used one within the own scientific discipline. Yet, they have to rely on the chosen method to properly address conceptual uncertainty and have to trust the obtained model rating - often without fully knowing under which premisses the models are rated.

In order to establish guidance in this field, I want to invert this pattern by answering the following research questions (RQ):

1. How is the law of *parsimony* implemented in different model rating and selection methods and what does this tell us about the evaluated models?

2. How are related multi-model frameworks properly used in specific *modelling settings* and how are their often contradictory results interpreted correctly?

3. How can a multi-model framework be chosen for any *modelling task* at hand such that the chosen framework properly addresses conceptual uncertainty specifically for this task?

These three research questions focus on different but complementary parts of the central question from Section 1.1: "Which multi-model framework employs the adequate Occam's razor with respect to the model setting of the modelling task at hand?"

To answer the first RQ, I collected various model rating and selection methods that were developed and used over the last decades and dissected them with respect to their interpretation of model complexity, i.e., their implementation of Occam's razor. Thereby, I merged insights about model selection from vastly different scientific disciplines and transferred them to applied modelling. As result, I developed a classification scheme for model selection criteria that allows to rate models according to whether a true model of an underlying DGP shall either be identified or only approached in an either Bayesian or non-Bayesian way. For each class, I discuss examples from hydrosystem modelling and propose matchings between certain modelling goals and model selection methods.

To answer the second RQ, I applied three Bayesian multi-model frameworks (BMA, Pseudo-BMA and Bayesian Stacking) to a finite model set for a typical task in hydrosystem modelling. Using the insights about the Occam's razor implementations in my classification scheme from RQ 1, I analysed and contrasted how the three frameworks account for conceptual uncertainty in philosophically different model settings (see (a), (b) and (c) in Section 1.1). Therefore, additionally, I propose a new practical model setting (called Quasi-$\mathcal{M}$-closed) to close a gap between the

existing ones for applied modelling. To assure reliability of the model rating results within the frameworks, I further applied the so-called Bayesian Bootstrap method. This method allows to address insufficiently confident model ratings of predictive performance in case of using only a small amount of available data. By this practical example, I link the rigorous evaluation of Bayesian multi-model frameworks with specified model settings and the usage of Bayesian Bootstrapping in the context of hydrosystem modelling. Thereby, I demonstrate how the Bayesian multi-model frameworks are adequately employed to achieve process-identification or predictive reliability.

To answer the third RQ, I clarify which Bayesian multi-model frameworks allow for model combination for either identifying or approaching a true model - similarly to the different kinds of model selection from RQ 1. Therefore, I introduce a recent approach that merges these principles by selecting the best model combination for process-identification and exemplify it for a potential hydrosystem modelling task. In direct relation to the three Bayesian multi-model frameworks from RQ 2, I propose a guiding scheme that allows to find the appropriate multi-model framework for an arbitrary modelling task at hand. The guide shows that choosing a proper multi-modelling framework is the natural outcome when starting from the philosophical perspective on the modelling task, rather than picking one based on some ad-hoc preference.

1.2.3 Thesis Structure

This thesis is structured as follows: First, I introduce the underlying theory on models and model settings, as well as state-of-the-art Bayesian multi-model inference methods in Chapter 2. Second, in three core chapters, I answer and discuss the three posed research questions:

- In Chapter 3, I answer RQ 1 and elucidate why some model selection criteria look similar but pursue vastly different goals of modelling. Thereby, the keyword is *model complexity*.
- In Chapter 4, I answer RQ 2 and demonstrate why respective model rating results seem to mean the same but represent vastly different takes on conceptual uncertainty. Thereby, the keyword is *model setting*.
- In Chapter 5, I answer RQ 3 and elaborate why some multi-model frameworks have similar names but apply to vastly different modelling situations. Thereby, the keyword is *model task*.

Third and finally, I deduce conclusions from the found answers and discuss potential issues for future research in Chapter 6.

2 Theory & Methods

Systematically addressing conceptual uncertainty requires a thorough consideration of model theory. Hence, I focus on model conceptuality and model settings in Section 2.1, on the principles behind Bayesian inference and uncertainty quantification in Section 2.2, on available Bayesian multi-model frameworks in Section 2.3, on their practical implementation in Section 2.4 and on common approximative methods for model rating in Section 2.5.

2.1 Model Theory

2.1.1 A Refined Model Definition

The previous chapter introduced two definitions of a model as "a thing used as an example to follow or imitate" or "a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions". Both are kept as general definition that, in the spirit of Cartwright (1983), considers a model as a tool to translate a set of hypotheses and/or theories into predictions (e.g. Nearing et al., 2016). In order to translate this into mathematical terms, we define a model here as: mathematical function of interrelated parameters $\boldsymbol{\Theta}$ that map a certain input $\boldsymbol{x}$ to an output $\boldsymbol{y}$ while being subject to noise/error $\boldsymbol{\epsilon}$. The model parameters comprise latent variables $\boldsymbol{\omega}$ like system properties and the "model frame" of boundary and initial conditions Γ ($\boldsymbol{\Theta} = \{\boldsymbol{\omega}, \Gamma_{\text{bound}}, \Gamma_{\text{init}}\}$):

$$M_m : \boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\Theta}) + \boldsymbol{\epsilon} \tag{1}$$

The above formulation for a model can most easily be understood as deterministic model, where all model parameters $\boldsymbol{\Theta}$ are fixed. Then, a certain input $\hat{\boldsymbol{x}}$ generates one specific $f(\hat{\boldsymbol{x}})$ which equals the model output if $\boldsymbol{\epsilon}$ is zero. Probabilistic models or deterministic models that are operated in a "stochastic framework" account for uncertainty of the components, e.g., specified by a probability distribution or probability density function (pdf) over parameters $p(\boldsymbol{\Theta})$, input $p(\boldsymbol{x})$ and noise $p(\boldsymbol{\epsilon})$. This results in a pdf of the model output $p(\boldsymbol{y})$. Therefore, stochastic modelling can be considered as generalization because taking the expectation over the assigned distributions yields a deterministic model.

A model set or model ensemble $\boldsymbol{M}$ refers to a finite list of N_M models that are fully specified and share at least one identical objective or QoI as output variable: $\boldsymbol{M} \coloneqq (M_1, M_2, ..., M_{N_M})$.

Despite its simplicity, the above definition of a model M_m as a thing to *follow* or *imitate* essentially spans the entire spectrum of model types when we specify what *follow* or *imitate* mean in the extremes (see, e.g., Breiman, 2001):

- *follow* refers to mechanistic modelling, where causal relations represent the modelled system. Mechanistic models help to understand and explain a system and allow for predictions based on causality. Universal natural laws and principles are specified for a certain system in a top-down manner. Physics-based models are an obvious example for mechanistic models and are often used synonymously. In hydro-system modelling, such a model would be the solution for the hydraulic head h as a function in space and time to the underlying parabolic partial differential equation (PDE) for groundwater flow:

$$S_0 \frac{\partial h}{\partial t} - \nabla(\boldsymbol{K} \nabla h) = Q_{in/out} \tag{2}$$

  with time t and ∇ as vector operator of partial derivatives in all spatial dimensions, parameters storage coefficient S_0 and hydraulic conductivity tensor $\boldsymbol{K}$, state variable hydraulic head h, and sources and sinks term $Q_{in/out}$ as boundary condition. Further, to specifically solve this mechanistic model, initial conditions like $h_0 = h(t_0)$ and boundary conditions like constant head or flux in space are required. Under steady-state conditions, the storage term in the groundwater flow equation drops out, turning the parabolic into an elliptic PDE $-\nabla(\boldsymbol{K} \nabla h) = Q_{in/out}$. Both contain the most fundamental law in hydrogeology as example for the physical flux-gradient relationship, i.e., Darcy's law for specific discharge $q = -\boldsymbol{K} \nabla h$.

- *imitate* refers to data-driven modelling, where associations of system variables are not necessarily causal. Data-driven models mimic a system and allow for predictions based on association of variables like correlation, regardless of whether there is causality or not. Patterns in the data are generalized for the observed or a similar system, i.e., a bottom-up approach. Empirical relations are exemplary for data-driven models, like a neural network (NN):

$$f(\boldsymbol{x}) = \psi(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{b}_0 \tag{3}$$

  A basic NN consists of interconnected nodes that allow for a non-linear mapping. Mathematically, in its core, it is a simple linear matrix multiplication

of parameters known as weights $\boldsymbol{W}$ between connected nodes and the vector of independent input variables $\boldsymbol{x}$. A linear offset parameter $\boldsymbol{b}$ represents the bias at each node and $\boldsymbol{b}_0$ additionally adjusts for the bias of the NN output. The so-called activation function $\psi(\cdot)$ (typically a sigmoidal function) introduces non-linearity and allows the NN with its linear core to fit also non-linear data. A typical example from hydrosystem modelling is the prediction of non-linear stream discharge as output based on inputs like rainfall, evaporation, stored water, etc.

2.1.2 Model Conceptuality: Model Types and Model Fidelity

The above section outlined two extremes of model types: *mechanistic and data-driven* modelling. These extremes are referred to as white-box or black-box models, respectively (Breiman, 2001). All mathematical models can conceptually be situated somewhere on the grey scale between or at these two extremes (Refsgaard and Abbott, 1996). Causality and association (specifically correlation) are not mutually exclusive - association at the white end is included into the model by causal relations and at the black end regardless thereof. A grey-box model is a mixture that uses causal relations for the parts of the model that are understood mechanistically and adds associative relations to approximate the system behaviour in other model parts. I refer to the black-white scale as model type in Figure 1, as explained in the following.

Just like the grey scale is continuous, the transitions between the model types are smooth. In natural science and engineering, systems and their processes are often represented by conservation laws, as for mass, energy and momentum. Corresponding white-box models contain ordinary or partial differential equations (see Equation 2) which fully describe causal relations for all involved variables and parameters. However, it is often not possible to describe every detail of a system by fully resolved physics at all scales, e.g., friction as a meso-scale phenomenon. Then, empirical or data-driven relations have to be employed for which it might be possible to assign physical dimensions as units (even with fractal exponents as in case of friction laws). Yet, they lump properties of the considered system that are not fully resolved or understood, e.g., by friction coefficients. The data-driven paradigm can likewise be used to model the whole system directly as full black-box model (see Equation 3) without the need of any mechanistic description. Illustrative examples of white-box, grey-box and black-box models in hydrosystem modelling with the same objective like stream discharge prediction are shown in Figure 1: finite element models (e.g. von Gunten et al., 2014), conceptual hydrologic (bucket) models (e.g. Fenicia et al., 2016) or neural network models (Zhang and Zhao, 2012), respectively.

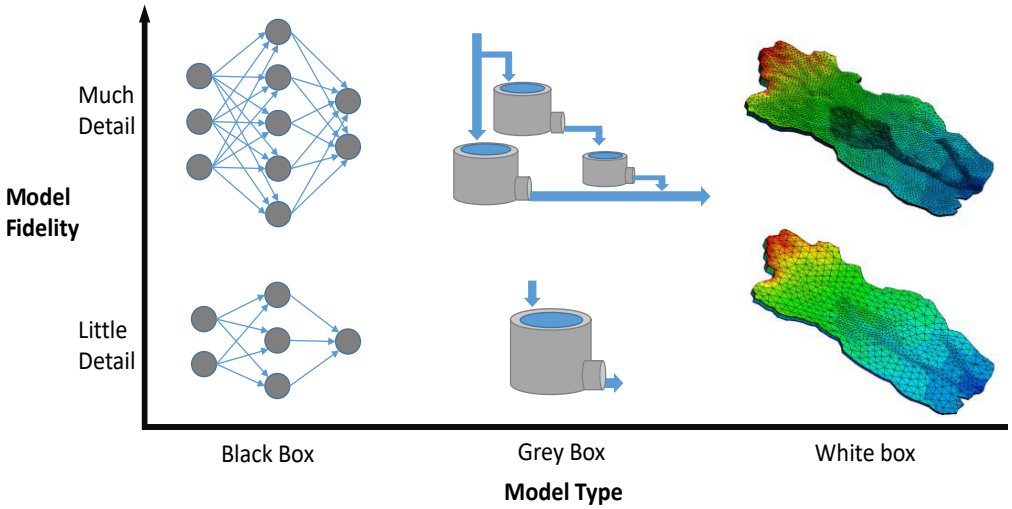


Figure 1: Model conceptuality: model type and fidelity; illustrated by two examples for each model type with different degrees of detail, i.e. a neural network, a bucket-type model and a PDE-based model (from von Gunten et al., 2014) for stream discharge modelling.

In each conceptual modelling type - black, grey or white - there are different levels of fidelity (see Figure 1). Typically, the term fidelity expresses accurate reproducibility. In modelling, this can refer to the model output or the model itself in terms of approaching or representing the true system. Here, model fidelity (e.g. Sinsbeck and Tartakovsky, 2015) refers to the number of model elements and the degree of detail or resolution within a particular model type - for which *fidelity* provides an encompassing and type-independent terminology. In case of black-box neural networks, fidelity refers to the number of nodes within each layer, the number of hidden layers as well as inter- and intra-linking connections, etc. Staying with the example of discharge predictions, high fidelity can mean to have two separate output nodes for base and peak flow. A grey-box model can have different fidelity stages depending on the amount of sequentially or parallely arranged units or pools (like hydrologic buckets), attached source and sink terms or other functional parts that modify the output like stream discharge. For differential equation-based white-box models like finite element models, the fidelity refers to different levels of spatial or temporal discretization (e.g. Leube et al., 2013), i.e. the grid resolution for solving the underlying equations, coupling of subsurface and surface water flow and further (semi-)physical relationships that refine process descriptions.

Model evaluation and rating usually does not only refer to the model conceptuality, but also to model fidelity. With respect to the scale of the modelling task (spatially, temporally, amount of data, etc.) and the resolution of available model parameters, boundary conditions, input variables and targeted output, the appropriateness of models varies along both categorical axes. Conceptual uncertainty refers to

the degree of appropriateness and, for compactness, is assumed to comprise both categories: model type and model fidelity.

2.1.3 Mathematical Spaces covered by Modelling

Regardless of which model classification system is used, every mathematical model M_m refers to three different spaces:

- The *prediction space* $\mathcal{Y}$ (a.k.a. data or model output space) contains our quantity of interest (QoI), e.g., hydraulic head or stream discharge. The variables $\boldsymbol{y}$ for model output (predictions) and $\boldsymbol{D}$ for observations (data) of the QoI are situated in $\mathcal{Y}$. The dimensions of this space have the units of the QoI. Each observation, i.e., every element in $\boldsymbol{D}$ or new data point D_o, adds an axis to this space. Models are set up to to match all observations by their output $\boldsymbol{y}$ or y_o on the respective axis. Note, that apart from predictions to meet observed data, prediction might also refer to not (yet) observed states or additional QoI. These can be considered as additional dimensions of $\mathcal{Y}$ yet without measured D_o. Model output is then supposed to yield estimates for both, measured and unmeasured quantities. However, for quantitative model rating only the dimensions of $\mathcal{Y}$ where data is available are considered. Meanwhile, the remaining dimensions provide more information on the model's plausibility in a qualitative and semi-quantitative manner, e.g., if the magnitude of the estimated but unobserved quantity matches the modeller's educated guess.

- The *parameter space* Ω_m encloses the parameters $\boldsymbol{\Theta}_m$ of a certain model, e.g., hydraulic conductivity or storage coefficients. Each model M_m has its own parameter space with parameters $\boldsymbol{\Theta}_m$. The dimensionality of the parameter space is defined by the number of model parameters $N_{p,m}$, which is sometimes referred to as parametric complexity (Vanpaemel, 2009).

- The *model space* $\mathcal{M}$ is populated by our model(s), e.g., physics-based or bucket-type models. Conceptually, each model M_m can be located somewhere in this space. However, the dimensions of this model space are not clearly defined. One can think of the dimensions that span this space as: number of model parameters, degree of nonlinearity of functional relations in the model, etc. Nonetheless, there is no comprehensive and complete description of $\mathcal{M}$.

In stochastic modelling, probability distribution are assigned to both, $\mathcal{Y}$ and Ω, in order to account for predictive and parametric uncertainty, respectively. Yet, without clear definition of $\mathcal{M}$, the assignment of a probability distribution the model space is not straight-forward without further specification.

2.1.4 Model Space Settings

All members in a model ensemble $\boldsymbol{M}$ (e.g., candidates from Figure 1) can be located somewhere in $\mathcal{M}$. Therefore, despite being only vaguely defined, we need a conceptual basis for model evaluation and comparison in $\mathcal{M}$ under the finite hypotheses problem (see Chapter 1; Nearing et al. (2016)), i.e. a qualitative system to formally relate the models in $\boldsymbol{M}$ to the data-generating truth a.k.a. true model M_{true}. The true model is the exact mathematical description of the system to be modelled, and is often also called the data-generating model or process, respectively (DGM or DGP). The above terms are used synonymously in the following. All observations $\boldsymbol{D}$ are per definition instances of the corresponding distribution of data / predictions from the true model $q(\boldsymbol{y}|M_{\text{true}})$. The way model candidates relate to M_{true} for a certain modelling task at hand can be distinguished by three different $\mathcal{M}$-settings adopted from Bernardo and Smith (1994):

- $\mathcal{M}$-closed: One of the models in the ensemble $\boldsymbol{M}$ is exactly the true model. Yet, it is unknown which one.
- $\mathcal{M}$-complete: None of the ensemble members M_m is the true model. The true model exists but it has not been possible (yet) to fully formulate it. Although no member fully represents the truth, at least one might still approximate it.
- $\mathcal{M}$-open: None of the ensemble members M_m is the true model; it is questionable whether a tractable true model exists or certain that it does not. Opposed to the other settings, the true model cannot even be conceptually defined due to, e.g., lack of expertise, lack of time, difficulty in conceptualizing, or the system is indeed infinitely complex.

All settings are visualized in Figure 2 as a projection of all model candidates from the $\mathcal{M}$-space onto a 2-dimensional plane (similar to e.g. Sanderson et al., 2015): Between each model's predictive distribution $p(\boldsymbol{y}|M_m)$ and the DGP's distribution $q(\boldsymbol{y}|M_{\text{true}})$, distances are evaluated using a statistical distance metric (cf. Section 2.2.4). Then, all models are projected on a 2D plane in a so-called multidimensional scaling process that preserves these mutual distances. Note, that this process has no unique solution regarding the allocation of models on the plane (Sanderson et al., 2015), but this does not limit its suitability for a schematic visualization.

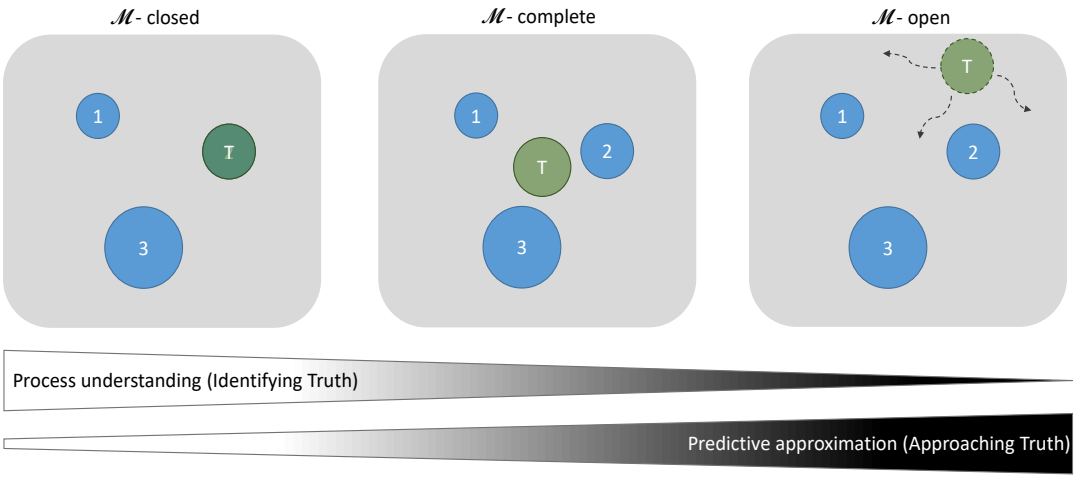


Figure 2: Illustration of the three $\mathcal{M}$-settings as 2D projection: $\mathcal{M}$-closed (left), $\mathcal{M}$-complete (center) and $\mathcal{M}$-open (right). The model set comprises three models (blue circles) of different complexity (indicated by the circle size). While in the $\mathcal{M}$-closed and $\mathcal{M}$-complete setting the true model (green circle with "T") is static in the model space, arrows in the $\mathcal{M}$-open setting depict the true model as "moving target". The primary objective (process-understanding or predictive approximation) in each setting is visualized by the grey scale (bottom).

In Figure 2, each circle can be considered as the outline of a model's projection on this plane. The calculated distances between the models can be found between the centers of the circles and the size of each circle sketches the complexity of the model. The transparent green circle resembles the true model and the enumerated opaque blue circles 1, 2 and 3 are model alternatives that are set up to *follow* or *imitate* this truth. Regarding (continuously) taken observations from the true model, Figure 2 can be read as follows:

- In the $\mathcal{M}$-closed setting, one of the models matches the true model exactly which follows from the fact that the DGP can be and is fully conceptualized and also fully formulated. Informative observations allow to identify one model in the set as the true model.

- In the $\mathcal{M}$-complete setting, it can only be incompletely formulated despite full conceptualization. Hence, the true model is not matched by any single model in the ensemble but it is known to be fixed and finite somewhere in $\mathcal{M}$. Informative observations allow to locate the true model with respect to the models in the set.

- In the $\mathcal{M}$-open setting, the truth cannot even be conceptualized, let alone written down. Then, there is no way to match the truth since the truth itself could not even be located statically on the 2D plane - it "moves" along (yet) unknown or hidden dimensions of $\mathcal{M}$. Informative observations allow to

reveal (previously unknown) features of the true model but without locating it.

These qualitative differences of the $\mathcal{M}$-settings are summarized in Table 1.

Table 1: Qualitative summary of the three $\mathcal{M}$-settings: $\mathcal{M}$-closed, $\mathcal{M}$-complete and $\mathcal{M}$-open with respect to the true model.

Model (pdf)...	$\mathcal{M}$-closed	$\mathcal{M}$-complete	$\mathcal{M}$-open
... can be conceptualized	fully	fully	incompletely
... can be formulated	fully	incompletely	impossibly
... matches actual true model (pdf)	fully	maybe closely	maybe temporarily

When referring to the basic purposes for modelling, i.e., to *follow* or "process understanding" and to *imitate* or "predictive approximation", we can simply visualize these $\mathcal{M}$-settings on a white-to-black scale as in Figure 2: The white end refers to $\mathcal{M}$-closed, while the black end resembles $\mathcal{M}$-open and the grey area in between contains $\mathcal{M}$-complete.

Each end has one *dominant objective*: At the white end, the goal can be to fully explain the DGP - via identifying the true model from our ensemble of models. At the black end, the objective can only be predictive capability - via selecting one or combining several of the models in the ensemble for obtaining best predictions. This does not mean that the respective other objective is discarded, but every multi-model framework is primarily tailored to accomplish one major objective, depending on what can be achieved in a certain $\mathcal{M}$-setting.

Although pursuing only *one* of the primary objectives, any multi-model framework might thereby still achieve the respective other objective: The correctly identified DGP in the $\mathcal{M}$-closed setting will automatically yield best predictions. Vice versa, the best model (combination) that produces best predictions outside of $\mathcal{M}$-closed might reveal variable associations or functional relations that are the reason for such predictive power. Potentially, these can be translated into a mathematical description that might help to (partially) understand the DGP - even if we know that at the black end, we are not able to fully conceptualize (and write down) the true model. In both cases, the respectively other objective is covered as a side-product while pursuing the major objective.

Coming from the perspective of physical science and engineering, the colors black and white in the extremes directly resemble the respective model categories that we think are able to fulfil the purpose of modelling in the specific $\mathcal{M}$-setting:

- *White-box models* that fully describe the causal relations of a system (like

physical differential equations) are the closest resemblance of a real-world DGP and therefore fit to the $\mathcal{M}$-closed setting (white end).

- *Black-box models* are assumed not to contain any physics and are therefore perfectly suited for the $\mathcal{M}$-open setting. There, we expect that the true DGP cannot even be conceptualized and a bottom-up (data-driven) approach for generalization is required at the black end.

The famous "all models are wrong, but some are useful" (Box, 1976) holds outside of the $\mathcal{M}$-closed setting (with increasing severity towards $\mathcal{M}$-open). Usually, when the word "model" is used, it is implicitly assumed that the modelling task at hand is outside of $\mathcal{M}$-closed - hence the quote is so appealing. However, in a scenario where an allegedly true model of the DGP is formulated and becomes part of the ensemble for process identification, the quote does not hold. A simple example for a true model can be found in the field of electromagnetism. There, the Maxwell-equations provide a true model of electromagnetic phenomena. Hence, under the current state of knowledge about physics, they are considered right and because of this, they are useful as a model.

It is important to internalize what statements can and cannot be made ultimately when comparing models while being in one or the other $\mathcal{M}$-setting: In an actual $\mathcal{M}$-closed setting, the best model resembles the DGP. There, and only there, it can be called true model. Per definition, the true model is fully consistent with the data, it provides the exact explanation and yields best predictions. Yet, outside of this framework, the model that yields best predictions by no means also resembles the actual DGP - it might not even be close, e.g., when we have a true physical system and use a data-driven approach to successfully mimic it. Even if a model rating clearly shows one model in the ensemble to be superior to the alternatives in terms of predictive power and we think it resembles the truth quite well, we can never state that we found the true model being outside of the $\mathcal{M}$-closed setting. But it still is the objectively best model for predictive approximation of the truth.

The unresolvable issue is that we never know which setting applies to our modelling task at hand. However, to handle multiple models in a multi-model framework, this is also not necessary as long as we understand which $\mathcal{M}$-setting is assumed by the applied method. The distinction between the $\mathcal{M}$-settings helps us in two respects:

- To choose a multi-model framework that at least helps us to achieve our primary modelling goal, i.e., to *follow* (understand) or *imitate* (predict).
- To correctly interpret the outcome of multi-model frameworks and properly account for conceptual uncertainty.

2.2 Bayesian Inference and Uncertainty Quantification

2.2.1 Uncertainty, Knowledge and Bayesian Priors

Uncertainty in modelling arises from two sources: true randomness and lack of knowledge (e.g. Rinderknecht et al., 2012, and references therein):

- True randomness as it appears in quantum mechanics, for example as radioactive decay, is called *aleatoric* uncertainty.
- Lack of knowledge about the system that shall be modelled, its thorough conceptualization, the correct mathematical description and all their cascading consequences are referred to as *epistemic* uncertainty.

Both kinds of uncertainty can mathematically be handled by probabilities (Rinderknecht et al., 2012), yet with two different perspectives: Frequentist and Bayesian.

Frequentist: Classically, the entry point to probabilities is the quantification of true randomness by frequencies of occurences in an infinite number of repetitions (e.g. Omlin and Reichert, 1999), e.g., for radioactive radiation as result of nuclear decay. Thereby, it is impossible to predict when a certain atom will decay but between a lot of atoms, it is possible to count how many of them decay after a certain time. This so-called frequentist perspective naturally yields probabilities in the sense of aleatoric uncertainty. Ultimately, this kind of uncertainty remains in any physical system and can be fully described by frequency-based probabilities.

Bayesian: Yet, before aleatoric uncertainty dominates, lack of knowledge is the major source of uncertainty. This can hardly be captured by a frequentistic consideration of whether something occurs or not. Instead, it can be described by degree of belief in the available knowledge. This is the Bayesian interpretation of probabilities. This kind of belief shall not be confused with arbitrariness or claims that cannot stand scientific reasoning. Only knowledge that follows the scientific paradigm and cannot be clearly falsified by experts can be transferred into probabilities that express degree of belief with respect to epistemic uncertainty.

Under the Bayesian paradigm, probabilities can be assigned before any data is available as so-called *prior* knowledge. Likewise, the distribution of errors between observations and model predictions that stem from the measurement process can also be considered as belief (Omlin and Reichert, 1999). As most important property, the Bayesian interpretation of probabilities as beliefs allows their updating when new evidence like observed data is available. Although the Frequentist and Bayesian probability interpretations differ, "resulting knowledge quantifications will be consistent with the axiomatic foundation of probability theory" (Rin-

derknecht et al., 2012). This means that, ultimately, after theoretically all lack of knowledge has been removed, the remaining epistemic uncertainty equals the irreducible aleatoric uncertainty.

Note that, despite being commonly used, the distinction between aleatoric and epistemic uncertainty is also subject to discussion (Nearing and Gupta, 2018). Further, other approaches of uncertainty quantification are not mutually exclusive with probability theory, e.g., information-theoretic tools allow for a broader analysis of uncertainty (cf. Section 2.2.4) in probabilistic forecasting.

The perspective on probabilities as degree of beliefs requires a closer look at the kinds of knowledge or belief that are transferred into (Bayesian) probability distributions, i.e., a distinction between objective, subjective and intersubjective knowledge (Gillies, 2012; Rinderknecht et al., 2012; Omlin and Reichert, 1999):

- Objective knowledge refers to facts that can be empirically confirmed and also hold in the absence of human opinion.
- Subjective knowledge, in contrast, is based on impressions of individuals that might but do not have to coincide with objective facts.
- Intersubjective knowledge is what several individuals agree upon.

While subjective knowledge appears and disappears with an individual, intersubjective knowledge remains as long as individuals join and stay with what was agreed upon. From a scientific perspective, intersubjective knowledge of experts about not-man-made systems is supposed to converge towards objective facts.

Often, the Bayesian interpretation of probabilities is criticized for its lack of objectivity or directly blamed to be fully subjective (as discussed in, e.g., Gelman et al., 2008). Typically, there is no objective prior knowledge available - there is even an ongoing scientific debate about what a truly objective prior is supposed to look like (van der Linde, 2012; Gelman et al., 2008). In contrast, fully subjective beliefs usually oppose scientific neutrality. Hence, the kind of knowledge that allows to conduct Bayesian inference in compliance with scientific requirements is mostly intersubjective (Reichert et al., 2015). Intersubjectivity entails a natural self-correction away from individual perspectives due to averaging of underlying subjective knowledge and incorporation of available facts that can ideally be objectively confirmed. Hence, Bayesian priors should be assigned and confirmed by expert knowledge (O'Hagan et al., 2006). Despite all criticism, the strength of the Bayesian approach is that it forces the modeller to state all induced knowledge and made assumptions explicitly.

In Bayesian (model) inference, we need to internalize that, first, we primarily deal with epistemic uncertainty, and, second, we need to scrutinize prior beliefs and respectively assigned probabilities in light of their kind of knowledge. When testing basic physical laws in perfectly controlled and isolated laboratory experiments, the randomness can be objectively described and even be interpreted from a frequentist perspective. In applied fields (e.g., hydrosystem modelling), where we work with a predefined limited subset of models and hardly identifiable parameters, objectivity is per se restricted. At the same time, a model comparison that is based on only subjective assignments of priors can be considered to be manipulated (Gelman et al., 2004) and is therefore non-trustworthy. For a model comparison to be reliable, we have to make sure that we conduct and interpret it from an intersubjective perspective if objective knowledge is unavailable, honouring that uncertainty in model evaluation and comparison is primarily of epistemic nature.

2.2.2 Bayesian Model Inference

Bayesian statistics provide a coherent framework to conduct uncertainty quantification in terms of uncertain knowledge (about model input, output, parameters and the conceptuality itself) for stochastic models (Gelman et al., 2004). Bayes' theorem thereby is the tool for updating knowledge with respect to new evidence - formally, by conditioning marginal distributions $p(\cdot)$ in order to obtain conditional distributions of the form $p(\cdot|\cdot)$. Practically, this enables us to navigate between and link the probabilistic distributions on the three modelling spaces introduced in Section 2.1.3, or, more specifically, their contained variables $\boldsymbol{y}$ (model output/predictions), $\boldsymbol{D}$ (data), $\boldsymbol{\Theta}$ (parameters) and $\boldsymbol{M}$ (models). For any given model M_m, Bayes' theorem writes as:

$$p(\boldsymbol{\Theta}_m|\boldsymbol{D}, M_m) = \frac{p(\boldsymbol{D}|\boldsymbol{\Theta}_m, M_m)p(\boldsymbol{\Theta}_m|M_m)}{p(\boldsymbol{D}|M_m)} \tag{4}$$

The marginal distribution $p(\boldsymbol{\Theta}_m|M_m)$ represents the prior distribution of parameters, i.e., before observations $\boldsymbol{D}$ of the QoI are considered in model M_m. As discussed in Section 2.2.1, it is supposed to represent intersubjective or even objective knowledge. Including $\boldsymbol{D}$ yields the posterior parameter distribution $p(\boldsymbol{\Theta}_m|\boldsymbol{D}, M_m)$ which is the prior distribution conditioned on $\boldsymbol{D}$. The conditioning is conducted via the likelihood function $p(\boldsymbol{D}|\boldsymbol{\Theta}_m, M_m)$ (cf. Section 2.2.3). Figuratively, the likelihood "pulls" and "squeezes" the prior while updating it to the posterior as shown in Figure 3.

The denominator $p(\boldsymbol{D}|M_m)$ in Equation 4 is the marginal likelihood of model M_m, i.e. the likelihood function integrated over the whole prior parameter distribution:

$$p(\boldsymbol{D}|M_m) = \int p(\boldsymbol{D}|\boldsymbol{\Theta}_m, M_m)\, p(\boldsymbol{\Theta}_m|M_m)\, \mathrm{d}\boldsymbol{\Theta}_m \tag{5}$$

The marginal likelihood is a measure for the average likelihood that model M_m generated $\boldsymbol{D}$. Therefore it is often termed Bayesian model evidence (BME). BME is a specific instance of the prior predictive density $p(\boldsymbol{y}|M_m)$ that formally writes like Equation 5 but with the specific observations $\boldsymbol{D}$ substituted by the general QoI $\boldsymbol{y}$:

$$p(\boldsymbol{y}|M_m) = \int p(\boldsymbol{y}|\boldsymbol{\Theta}_m, M_m)\, p(\boldsymbol{\Theta}_m|M_m)\, \mathrm{d}\boldsymbol{\Theta}_m \tag{6}$$

Similarly, the conditional density $p(\boldsymbol{y}|\boldsymbol{\Theta}_m, M_m)$ (i.e., the likelihood for $\boldsymbol{y} = \boldsymbol{D}$) can also be marginalized over the parameter posterior which yields a posterior predictive distribution:

$$p(\boldsymbol{y}|\boldsymbol{D}, M_m) = \int p(\boldsymbol{y}|\boldsymbol{\Theta}_m, M_m)\, p(\boldsymbol{\Theta}_m|\boldsymbol{D}, M_m)\, \mathrm{d}\boldsymbol{\Theta}_m \tag{7}$$

Assuming "new" observations $\boldsymbol{D}'$ as specific instances from $q(\boldsymbol{y}|M_{\text{true}})$ and a posterior that stems from conditioning on "old" observations $\boldsymbol{D}$, equation 7 provides a posterior predictive density $\boldsymbol{D}'$ that formally looks similar to the BME in equation 5. However, the posterior-marginalized likelihood is an integrated measure of the predictive power under the updated parameter belief and does not tell us anything about having generated $\boldsymbol{D}$ (cf. Section 2.3).

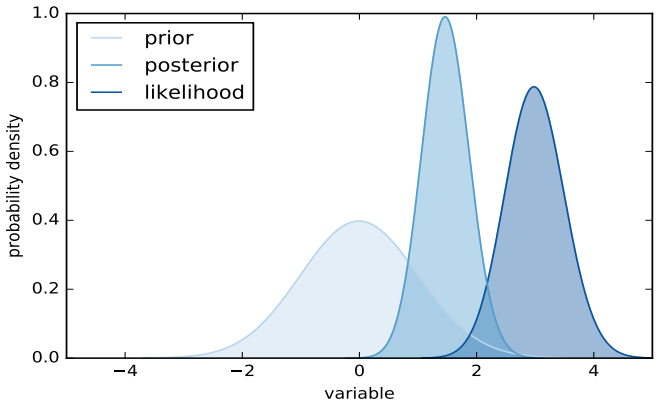


Figure 3: Illustration of Bayesian inference for a variable (like a parameter Θ or a prediction y) with a Gaussian prior pdf (light blue, left), with a Gaussian likelihood (dark blue, right) and the corresponding Gaussian posterior pdf (medium blue, center).

Bayes' Theorem can be used (Equations 4 to 7) regardless of which $\mathcal{M}$-setting applies. However, the meaning of its parts differs slightly between the settings:

- In the $\mathcal{M}$-closed setting, it is assumed that the true model is in the ensemble. Hence also the prior parameter distribution of this true model is assumed to be correct which means there is no systematic deviation between the model's parameter distribution and the true parameter distribution. The likelihood function therefore only fulfills the purpose to account for measurement error and thereby adjust the model output $\boldsymbol{y}$ to the "blurredness" of data $\boldsymbol{D}$. Figuratively, the likelihood function acts like glasses through which the model can adopt to the sharpness of the observations.

- Outside of $\mathcal{M}$-closed, a prior is an "educated guess" about the parameters for *per se* wrong models. It is assigned such that each model reaches highest predictive capability despite being conceptually wrong. Nonetheless, there is a systematic offset between all models and the truth and hence also for their parameter distributions (see Figure 3 for illustration) - if the models and their corresponding parameters are conceptually wrong, the assigned parameter prior cannot be right. Then, staying in the above imagery, in addition to adjusting for the resolution of observations, the likelihood function also might contain a "filter" like darkening in sun-glasses to adjust for the systematic deviation.

2.2.3 Likelihood Function and Prediction Errors

Bayesian updating from prior to posterior belief crucially depends on the choice of likelihood function. It defines how information about the system, that is contained in the observations, is transferred to the knowledge about model parameters $\boldsymbol{\Theta}_m$, yielding the posterior. Mathematically, the likelihood function resembles the assumed distribution of errors or residuals $\boldsymbol{r}$ between model predictions $\boldsymbol{y}$ and data $\boldsymbol{D}$. Originally, it is a frequentist approach (Fisher, 1922) but can likewise be used and interpreted under the Bayesian paradigm (Del Giudice et al., 2013).

In Bayes' theorem (Equation 4), the likelihood is $p(\boldsymbol{D}|\boldsymbol{\Theta}_m, M_m)$ which means that it is a function of $\boldsymbol{D}$ given $\boldsymbol{\Theta}_m$ of a certain M_m. However, when updating the parameter prior $p(\boldsymbol{\Theta}_m|M_m)$, we are interested in the dependence on $\boldsymbol{\Theta}_m$. Therefore, we employ the original model formulation in Equation 1 to make the likelihood a function of the model parameters $\boldsymbol{\Theta}_m$ given observed data $\boldsymbol{D}$: $\mathcal{L}(\boldsymbol{\Theta}_m|\boldsymbol{D})$. Note, that the re-labelled $\mathcal{L}(\boldsymbol{\Theta}_m|\boldsymbol{D})$ does not necessarily integrate to 1 and is therefore no proper probability distribution.

In most general terms, errors are often assumed to be independent and identically distributed (i.i.d.). A specific instance of this is Gaussian white noise, i.e., uncorrelated normally distributed errors with zero mean and finite variance. The corresponding likelihood function for N_s residuals for given data $\boldsymbol{D}$ is a multivariate Gaussian:

$$\mathcal{L}(\boldsymbol{\Theta}_m|\boldsymbol{D}) = (2\pi)^{-\frac{N_s}{2}}|\boldsymbol{R}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{r}(\boldsymbol{\Theta}_m)^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{r}(\boldsymbol{\Theta}_m)\right) \tag{8}$$

It describes the normal distribution of residuals $\boldsymbol{r}(\boldsymbol{\Theta}_m) = \boldsymbol{y}(\boldsymbol{\Theta}_m) - \boldsymbol{D}$, i.e., the distribution of predictions $\boldsymbol{y}(\boldsymbol{\Theta}_m)$ centred at the observed data $\boldsymbol{D}$. In case of uncorrelated errors, the variance-covariance matrix of errors $\boldsymbol{R}$ contains only diagonal elements. They represent the measurement uncertainty (see $\boldsymbol{\epsilon}$ in Equation 1) and can be interpreted as weighting factors for each residual - the larger the measurement uncertainty, the smaller the weight of the corresponding residual. Then, the exponential argument resembles the weighted sum of squared errors (WSSE), which makes the logarithmic likelihood proportional to the WSSE, i.e. a common error metric.

Depending on the modelling task at hand, alternative likelihood formulations might potentially be more suitable than Gaussian white noise. However, this normal error model allows easily to shed light on the issues involved in the adequate assignment of a likelihood function with respect to the $\mathcal{M}$-setting of the modelling task. Note, that the "Gaussian white-noise likelihood model" in Equation 8 does not account for a systematic bias between model predictions and observations that occurs when wrong models are used. Such bias can be accounted for within the likelihood function by modifying $\boldsymbol{R}$ (e.g., by having non-zero off-diagonal elements) or a separate error model (e.g. Del Giudice et al., 2013). However, while outside of the $\mathcal{M}$-closed setting such a statistical error treatment might help to increase predictive performance, it contradicts the idea behind identifying a true model in the $\mathcal{M}$-closed case according to which no systematic bias exists. Hence, in $\mathcal{M}$-closed, the likelihood function should only account for measurement uncertainty. Philosophically, another perspective is to automatically account for errors by making the model stochastic instead of describing them by a likelihood function (Nearing et al., 2016). Yet, mathematically, corresponding equations of these two perspectives are equivalent.

Further, the above likelihood formulation does not depend on the state of model variables, i.e., it cannot represent errors relatively to the values themselves. A common example from hydrologic modelling is that the measurement uncertainty of stream discharge depends on the flow regime, e.g., errors are relative to the

magnitude of discharge under low, medium or high flow conditions. This problem of variance that changes depending on the magnitude of the QoI is known as heteroscedasticity (Sorooshian and Dracup, 1980). Possible solutions in modelling are to make the elements in $\boldsymbol{R}$ dependent on the magnitude of the measurements, i.e., to assign relative errors, or to apply a transformation, e.g., the Box-Cox-Transformation (Box and Cox, 1964), to rescale the values and control the variance. This, however, might introduce additional (uncertain) transformation parameters and requires an adjustment of the likelihood function (see, e.g., Schöniger et al., 2014). Alternatively, the likelihood has to be evaluated over all possible states, which mathematically resembles an expensive integration that is often analytically intractable and computationally infeasible (Albert et al., 2015). Then, so-called Approximative Bayesian Computation (ABC) methods allow for sampling the likelihood function rather than fully evaluating it and this way to pursue Bayesian inference. Further, ABC methods allow to infer an approximate posterior using summary statistics of the QoI if the model output space $\mathcal{Y}$ is high-dimensional and a likelihood function like Equation 8 becomes unsuitable (Albert et al., 2015).

Opposed to this are approaches that do not employ a rigorous likelihood definition that would allow to infer a full probability distribution of model output $\boldsymbol{y}$. A respective popular method in hydrology is the so-called GLUE (generalized likelihood uncertainty estimation; Beven and Binley, 1992). It uses a certain rescaled error metric (related to "acceptable" not to probable errors) to weigh model predictions. Based on this, prediction envelopes are delimited and the whole model is rated in its forecast performance for comparison against alternatives. Note, that while such an approach provides pragmatic estimates of acceptable predictions and corresponding ranges of variability, they should rather be considered as weighted sensitivity analysis (Montanari, 2007) and do not allow for rigorous probabilistic uncertainty quantification from a Bayesian perspective.

2.2.4 Scores and Divergences

In probabilistic forecasting - as with prior or posterior predictive densities from Bayesian inference - so-called scoring rules are used to evaluate the quality of forecasts from a stochastic model over the entirety of the distribution (Gneiting and Raftery, 2007). For deterministic predictions, this is trivially done by calculating a certain error metric like the sum of squared errors (SSE) or the mean absolute error (MAE) between the deterministic prediction and the observed values (as it is also done when evaluating the likelihood function for specific parameter values). Thereby, the kind of error metric regulates the impact of each residual, e.g., taking squares amplifies the impact of large residuals. Model performance evaluation and

comparison is therefore always subject to the chosen metric. Regarding entire distributions of forecasts, scoring rules play the same role. Therefore, in the following, the so-called log-score, the Kullback-Leibler divergence, Shannon entropy and Brier score will be introduced.

The most popular scoring function in Bayesian inference is the *log-score* (Good, 1952), i.e., the logarithmic value of the forecast density $p(\cdot)$ for a random variable $\boldsymbol{Y}$, evaluated by inserting given data $\boldsymbol{D}$:

$$f_{\text{score}}(p, \boldsymbol{D}) = \ln(p(\boldsymbol{D})) \tag{9}$$

Every scoring rule has a corresponding divergence that resembles a "distance metric" between the predictive distribution of the model $p(\boldsymbol{y})$ and the target distribution $q(\boldsymbol{y})$, here, the true distribution $q(\boldsymbol{y}|M_{\text{true}})$. The log-score corresponds to the *Kullback-Leibler divergence*:

$$\mathrm{D}_{\mathrm{KL}}(q, p) = \int q(\boldsymbol{y}) \ln\left(\frac{q(\boldsymbol{y})}{p(\boldsymbol{y})}\right) \mathrm{d}\boldsymbol{y} = \mathrm{E}_q[\ln(q(\boldsymbol{y}))] - \mathrm{E}_q[\ln(p(\boldsymbol{y}))] \tag{10}$$

The Kullback-Leibler divergence (D_{KL}) is a metric for comparing two distributions that grows with the lack of their matching. It originates from information theory where it is also known as relative entropy (Weijs et al., 2010). Accordingly, the divergence between two distributions can also be interpreted from an information-theoretic perspective as information gain when moving from one distribution towards the other. An information gain of zero thereby implies perfect matching since the same information is contained in both distributions. Note, that moving from either one or the other distribution implies asymmetry, which formally disqualifies D_{KL} as distance metric although it carries a notion of how close distributions are to one another.

Correspondingly, the ties of the logarithmic score to information theory are obvious when looking at the fundamental information-theoretic concept of *Shannon entropy* as expected information per event: $H(p(\boldsymbol{y})) = \mathrm{E}_p[-\ln(p(\boldsymbol{y}))]$ (Shannon, 1948). Hence, the negative log-score is commonly interpreted as information-theoretic "surprisal" (Tribus, 1961). From this perspective, maximizing the log-score corresponds to minimizing surprisal which resembles high matching of distributions in a D_{KL} sense - no surprisal in predictions means perfect forecasts under uncertainty, i.e. perfect match of $p(\boldsymbol{y})$ and $q(\boldsymbol{y}|M_{\text{true}})$.

There are many other scores that rate probabilistic forecasts with different emphasis, e.g., the well-known quadratic a.k.a. *Brier score* with the L2-norm (Euclidian

distance) as associated divergence. However, the log-score comes with several beneficial properties (Boero et al., 2011), making it the only so-called proper local scoring rule (Gneiting and Raftery, 2007). A thorough discussion of scoring rules and divergences is beyond the scope of this thesis, but the interested reader may refer to Gneiting and Raftery (2007). For the evaluation of predictive distributions in Bayesian model inference the log-score with the corresponding D_{KL} is the preferred choice and used throughout this thesis.

In Bayesian inference, the model prediction performance is expressed as marginalized likelihood (Equations 5 to 7). Hence, its logarithm can be considered as an expected log-score and interpreted in a D_{KL} sense with respect to the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$. Often, for pragmatic reasons like the values' order of magnitude, the log-likelihood is already employed in Bayesian inference. Yet, the theoretical underpinnings of scoring rules puts the performance evaluation of Bayesian predictive distributions on a strict basis.

2.3 Bayesian Multi-Model Frameworks

In all Bayesian multi-modelling approaches, the common posterior predictive distribution $p(\boldsymbol{y}|\boldsymbol{D})$ of N_M models is:

$$p(\boldsymbol{y}|\boldsymbol{D}) = \sum_{m=1}^{N_M} p(\boldsymbol{y}|\boldsymbol{D}, M_m) w(M_m|\boldsymbol{D}) \tag{11}$$

where $p(\boldsymbol{y}|\boldsymbol{D}, M_m)$ is the posterior predictive distribution of $\boldsymbol{y}$ given data $\boldsymbol{D}$ only of model M_m and $w(M_m|\boldsymbol{D})$ is the posterior model weight of model $\mathcal{M}$ given $\boldsymbol{D}$. The model weights are simplex weights ($w_m \geq 0$ and $\sum_{m=1}^{N_M} w_m = 1$) in order to assure integration to one.

Correspondingly, the ensemble expectation $\mathrm{E}[\boldsymbol{y}|\boldsymbol{D}]$ is the weighted average of all model-specific expectations:

$$\mathrm{E}[\boldsymbol{y}|\boldsymbol{D}] = \sum_{m=1}^{N_M} \mathrm{E}[\boldsymbol{y}|\boldsymbol{D}, M_m] w(M_m|\boldsymbol{D}) \tag{12}$$

The variance $\mathrm{Var}[\boldsymbol{y}|\boldsymbol{D}]$ of the common posterior predictive distribution writes as:

$$\mathrm{Var}[\boldsymbol{y}|\boldsymbol{D}] = \sum_{m=1}^{N_M} \mathrm{Var}[\boldsymbol{y}|\boldsymbol{D}, M_m] w(M_m|\boldsymbol{D}) + \sum_{m=1}^{N_M} \left(\mathrm{E}[\boldsymbol{y}|\boldsymbol{D}, M_m] - \mathrm{E}[\boldsymbol{y}|\boldsymbol{D}]\right)^2 w(M_m|\boldsymbol{D}) \quad (13)$$

These "Bayesian Multi-Modelling" equations are the same for all considered Bayesian model selection, averaging and combination frameworks (see, e.g., Hoeting et al., 1999). Each approach is a certain kind of linear probability density function (pdf) averaging - in case all individual predictive pdfs were Gaussian, one obtains a Gaussian mixture as common distribution. Referring back to the 2D-plane projection of models in Figure 2 this means that the models stay where they are on the plane. The weights then define the fraction of each model's contribution to the common posterior predictive distribution $p(\boldsymbol{y}|\boldsymbol{D})$ in order to *follow or imitate* $q(\boldsymbol{y}|M_{\text{true}})$. Yet, the multi-model frameworks vastly differ in how each one estimates the model weights $w(M_m|\boldsymbol{D})$ and in terms of their meaning.

2.3.1 Bayesian Model Selection and Averaging

Model Selection: *Bayesian model selection* (BMS; e.g., Hoeting et al., 1999; Schöniger, 2016) is probably the most popular Bayesian multi-model framework. It originates from applying Bayes' Theorem not only to the parameters of a model but also to all model alternatives in a finite set of models, i.e., N_M fully specified model competitors. This is sometimes called the "two levels of inference" (MacKay, 1992). In the $\mathcal{M}$-closed setting, BMS is able to identify the true model among these ensemble members - which is the only logical goal. There, and only there, it is possible to assign a *probability to be the true model* to each ensemble member. Hence, model weights (see Section 2.3), both prior $w(M_m)$ and posterior $w(M_m|\boldsymbol{D})$, resemble probabilities. The mere fact that these probabilities sum up to one emphasizes that the true model *must* be in the considered set.

Prior model weights $w(M_m)$ are defined by the distribution that the modeller assigns to the list of models in the ensemble. For example, a discrete uniform distribution over N_M models would yield a model weight of $w(M_m) = p(M_m) = 1/N_M$ for each model alternative. Posterior model weights are updated model probabilities in light of evidence (data $\boldsymbol{D}$). They are gained by incorporating the model-specific marginal likelihoods (BMEs) $p(\mathbf{D}|M_m)$:

$$w_{\text{BME}}(M_m|\boldsymbol{D}) = p(M_m|\mathbf{D}) = \frac{p(\mathbf{D}|M_m)p(M_m)}{\sum_{k=1}^{N_M} p(\mathbf{D}|M_k)p(M_k)} \quad (14)$$

The marginal likelihoods are obtained by evaluating Equation 5 for each model. Ratios of marginal likelihoods between two models, BME_m and BME_k, are called Bayes factors (Kass and Raftery, 1995) and are a commonly employed tool to directly compare two models. Bayes factors (BF) multiplied with the ratio of the respective prior model probabilities yield posterior odds (e.g. Calderhead and Girolami, 2009) which further take into account the prior beliefs in the respective models:

$$\underbrace{\frac{p(M_m|\mathbf{D})}{p(M_k|\mathbf{D})}}_{posterior\,odds} = \underbrace{\frac{\text{BME}_m}{\text{BME}_k}}_{BF} \underbrace{\frac{p(M_m)}{p(M_k)}}_{prior\,odds} \tag{15}$$

With its implicit assumption of rating model in terms of their probability to be the true model, the BMS framework (in form of both, posterior model weights or posterior odds) naturally seeks to converge towards one model that it considers to be the correct one. This is called consistency in model selection (e.g. Hurvich and Tsai, 1989; Shibata, 1986). The preference for the allegedly true model thereby increases under growing data size. In the limit of infinitely informative data, BMS will asymptotically converge towards the true model in the $\mathcal{M}$-closed setting.

BMS shows direct ties to universal induction (see Chapter 1) which is also consistent in finding the data-generating truth. BMS is a practical framework to identify a true model under insufficient knowledge from a finite subset of model alternatives in face of uncertainty in all modelling spaces. Its relation to code-length based model evaluation and rating can be shown by decomposing BME as will be done and discussed in Section 2.5.

Model Averaging: Based on the obtained posterior model weights *Bayesian Model Averaging* (BMA; e.g., Hoeting et al., 1999) can be performed. Yet, due to the consistency property, the model weights from Equation 15 are not meant to yield weighted model averages in the sense of model combination as clearly stated by Minka (2002): "weights in BMA only reflect a statistical inability to distinguish the hypothesis [sic] based on limited data". Combination implies weights that converge to certain values such that the weighted average of models yields best predictions. Opposed to this, the weights in BMS represent the quantified conceptual uncertainty in converging towards the allegedly true model using only the currently available data $\boldsymbol{D}$. Averaging in this sense has to be interpreted as accounting for the conceptual uncertainty in selection when model predictions need to be made in an $\mathcal{M}$-closed setting despite incomplete identification of the true model (see Figure 4 and Höge et al. (2019)).

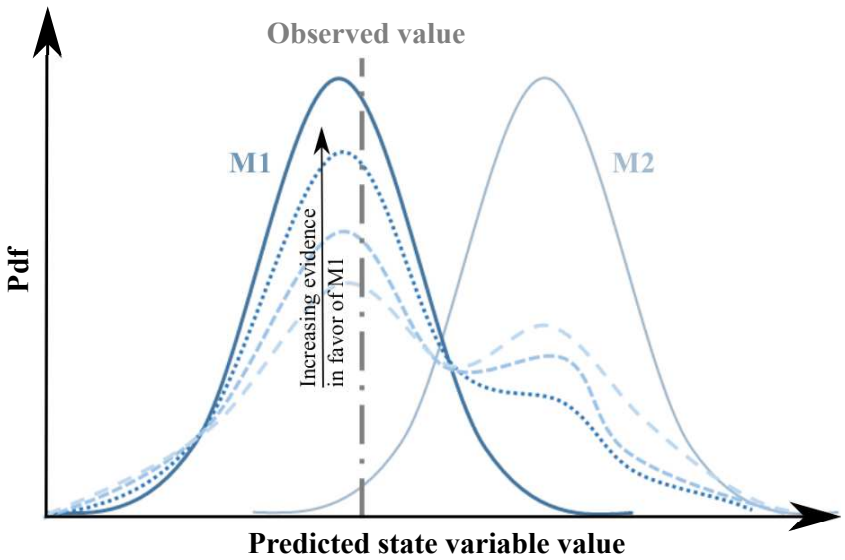


Figure 4: Predictive pdf resulting from applying BMA and BMS to two competing models M1 and M2. BMA predictive pdf (dashed lines) converges to BMS result (here: predictive pdf of M1, solid line) with increasing data set size (from Höge et al., 2019).

BMA is therefore an intermediate step throughout BMS for a snapshot of data $\boldsymbol{D}$. With increasing amount of data and subsequent re-iteration of Equation 14 on the model ensemble, conceptual uncertainty in selection decreases. The posterior model weights will change in favour of the allegedly true model until it ultimately receives a model weight of one and in $\mathcal{M}$-closed, $w(M_m|\boldsymbol{D}) = 1$ is an irrevocable statement of truth.

Interpretation: Consistency in model selection holds strictly only in an actual $\mathcal{M}$-closed setting. Being interpreted from the perspective of scores and divergences (cf. Section 2.2.4), BMA/BMS asymptotically selects the (allegedly true) model that minimizes the $\mathrm{D_{KL}}$ between its prior predictive pdf $p(\boldsymbol{y}, M_m)$ and $q(\boldsymbol{y}|M_{\mathrm{true}})$ (Yao et al., 2017) - it rates individual models, not their average. However, note that if applied outside of true $\mathcal{M}$-closed setting, this best model can still be arbitrarily "far" away from the the truth in an absolute sense but the closest relatively to its competitors. Outside of $\mathcal{M}$-closed, this might be mistaken for being the model that provides best predictions. A modeller who employs BMS/BMA outside of the the $\mathcal{M}$-closed setting risks to worsen explanatory power, predictive capability and reliable uncertainty quantification.

Regardless of the $\mathcal{M}$-setting in which BMS/BMA is employed, it is crucial to keep this information-theoretic interpretation in mind. BME is a particular instance of prior predictive density $p(\boldsymbol{D}, M_m)$ and rates unconditioned (uncalibrated) models

in their ability to have generated the observations. Interpreted as such, BME might therefore still be useful as so-called prior predictive check (Gabry et al., 2017) at $\boldsymbol{D}$, i.e. an assessment of whether the model candidate meets the pattern and magnitude of the data $\boldsymbol{D}$ in the way itself and its prior were set up. Yet, outside of the $\mathcal{M}$-closed setting, other multi-model frameworks than BMS/BMA, that are based on other quantities than BME, are more suitable to handle conceptual uncertainty (see, e.g., Section 2.3.2).

2.3.2 Predictive (Bayesian) Model Selection and Averaging

Model Selection: For a typical modelling task, the assumption of being in an actual $\mathcal{M}$-closed setting might rather be the exception than the rule. A Bayesian multi-model framework for selection (and averaging) that is applicable outside of the $\mathcal{M}$-closed setting is so-called predictive or Pseudo-BMS (Geisser and Eddy, 1979). It also employs Bayes' theorem, however not on the model level since model probabilities of being true cannot be assigned outside of the $\mathcal{M}$-closed setting. Pseudo-BMS does not seek to find a true model in the ensemble but the one with best predictive capability despite being wrong - the logical goal when a true model is unavailable. This is called non-consistent model selection. "Pseudo" in this context refers to two things:

1. The $\mathcal{M}$-setting of the modelling task at hand is assumed to be not "closed", which prohibits the interpretation of model weights as probabilities.

2. Formally, there is no big difference by how Pseudo-BMS is conducted compared to BMS/BMA.

Both rate models based on marginalized likelihoods. However, contrarily to the evaluation of BME (Equation 5) for BMS/BMA, Pseudo-BMS requires the likelihood of new unknown (out-of-sample) data $\boldsymbol{D'}$ marginalized over the posterior parameter distribution given (within-sample) data $\boldsymbol{D}$:

$$p(\boldsymbol{D'}|\boldsymbol{D}, M_m) = \int p(\boldsymbol{D'}|\boldsymbol{\Theta}_m) p(\boldsymbol{\Theta}_m|\boldsymbol{D}, M_m) \mathrm{d}\boldsymbol{\Theta}_m \tag{16}$$

$p(\boldsymbol{D'}|\boldsymbol{D}, M_m)$ is a specific instance of the general posterior predictive density $p(\boldsymbol{y}|\boldsymbol{D}, M_m)$ of the model (Equation 7) at $\boldsymbol{D'}$ (Gelman et al., 2014). Since we do not have access to future data $\boldsymbol{D'}$, we can split the available $\boldsymbol{D}$ into calibration and validation data. This is the idea behind cross-validation (CV; e.g., Stone, 1977): Adjusting the parameters by using the calibration part and rating the models using the validation part.

This way, in deterministic model selection, the expected predictive error (EPE) is estimated as basis for model rating. Doing this probabilistically is referred to as Bayesian Cross Validation (BCV) (Piironen and Vehtari, 2017) to estimate predictive densities. Thereby, leave-one-out (LOO) cross-validation is the closest approximation to actual future data by always holding out one data point D_o from data $\boldsymbol{D}$ and inferring on the rest of the data set $\boldsymbol{D}_{\varnothing}$. The pointwise posterior predictive density writes as:

$$p(D_o|\boldsymbol{D}_{\varnothing}, M_m) = \int p(D_o|\boldsymbol{\Theta}_m)p(\boldsymbol{\Theta}_m|\boldsymbol{D}_{\varnothing}, M_m)\mathrm{d}\boldsymbol{\Theta}_m \tag{17}$$

To resemble potential future data $\boldsymbol{D}'$, every data point $D_o \in \boldsymbol{D}$ is considered separately. Exploiting the therefore imposed i.i.d. assumption about all D_o, we take the product of the pointwise predictive densities $p(D_o|\boldsymbol{D}_{\varnothing}, M_m)$. On a logarithmic scale, this turns the product into a sum of expected logarithmic pointwise predictive densities (Gelman et al., 2014), i.e. the sum of expected pointwise log-scores:

$$elpd_{LOO,m} = \sum_{o=1}^{N_s} \ln p(D_o|\boldsymbol{D}_{\varnothing}, M_m) \tag{18}$$

The logarithmic predictive density is an expected value, because the predictive capability of a model is estimated for one data point D_o from $q(\boldsymbol{y}|M_{\text{true}})$. Since each data point is yet unknown, the estimation has to be formulated as expectation $\mathrm{E}_q[\cdot]$ over the whole $q(\boldsymbol{y}|M_{\text{true}})$. Finally, exponentiating $elpd_{LOO,m}$ yields an approximation to equation 16 and model weights via:

$$w_{\text{LOO}}(M_m|\boldsymbol{D}) = \frac{\exp(elpd_{LOO,m})}{\sum_{k=1}^{N_M} \exp(elpd_{LOO,k})} \tag{19}$$

These Pseudo-BMS weights (often also called (B)CV-type weights or Akaike weights; Yao et al., 2017) shall not be confused with BMS/BMA weights from equation 14. They cannot be interpreted as model probabilities, but provide a relative measure for how close each (wrong) model's posterior predictive density $p(\boldsymbol{y}|\boldsymbol{D}, M_m)$ is towards the true DGP distribution $q(\boldsymbol{y}|M_{\text{true}})$ given the current data $\boldsymbol{D}$. Although not resembling probabilities, the definition of weights assures $\sum_{m=1}^{N_M} w(M_m|\boldsymbol{D}) = 1$ as natural constraint in pdf-averaging (see Yao et al., 2017).

Plainly, Pseudo-BMS weights reflect how strongly a certain (wrong) model is supported by $\boldsymbol{D}$. Under assumed growing data size, the data $\boldsymbol{D}$ at each stage is used to assess how each model candidate, specifically its conceptuality and parameters, makes use of the information contained in the available limited $\boldsymbol{D}$ in order to

achieve highest predictive density for out-of-sample $\boldsymbol{D}'$. In model selection, this is sometimes referred to as asymptotic efficiency (Shibata, 1980). In the asymptotic limit of infinite data, the best predictive model out of the competitors within the ensemble will be elicited in an $\mathcal{M}$-complete/-open setting.

The Pseudo-BMS framework is the Bayesian equivalent to selecting models deterministically in terms of their validation performance as out-of-sample error. The similarity between the probabilistic and deterministic approach becomes evident in the comparison between approximative methods for estimating predictive density or predictive error (see Section 2.5).

Model Averaging: Pseudo-BMS can be most easily understood when compared to BMS/BMA in terms of their similarities and differences. The selective behaviour of Pseudo-BMS is similar to BMS/BMA, yet with different objective. Model weights can also be used to average the models but they do not yield an average in the sense of combination either. The weighted average in Pseudo-BMS/BMA honours the conceptual uncertainty between (wrong) models to be the best predictive one in light of the currently available data $\boldsymbol{D}$. Hence, the model weights of ensemble members also resemble only an intermediate stage on the way towards selection and are supposed to change under re-iteration of the framework with growing data (cf. Figure 4 interpreted for posterior predictive distributions).

Yet, Pseudo-BMS/BMA vastly differs from BMS/BMA, because the model set can change by adding new (or dropping old) model candidates throughout re-iteration (see Leeb and Pötscher, 2009). With the underlying assumption to be outside of the $\mathcal{M}$-closed setting, there is no requirement as in BMS/BMA to identify the true model out of a fixed and finite set with assigned distribution. In the Pseudo-BMS/BMA framework, the only objective for a model in the set is to provide a high predictive score. It can therefore be arbitrarily complex (or simple) as long as it is quantifiably supported by data $\boldsymbol{D}$. This argument can even be taken further, such that if a model ever receives a model weight of one, there might be enough data to support an even more complex model as best predictive model - in the $\mathcal{M}$-complete setting to approximate the unknown but "static" true model more closely and in the $\mathcal{M}$-open setting to chase the truth as "moving target".

Interpretation: The model with highest weight in Pseudo-BMS/BMA is relatively closest to the unknown truth in an information-theoretic sense, but can still be arbitrarily far away in absolute terms. Being interpreted from the perspective of scores and divergences (cf. Section 2.2.4), Pseudo-BMS/BMA asymptotically selects the model that minimizes the D_{KL} between its posterior predictive pdf

$p(\boldsymbol{y}|\boldsymbol{D}, M_m)$ and the unknown $q(\boldsymbol{y}|M_{\text{true}})$ (Yao et al., 2017). Again, this holds for the posterior predictive densities of individual models, not for their average.

This proximity can be shown for the $\mathcal{M}$-complete setting (Le et al., 2017) because the true model distribution $q(\boldsymbol{y}|M_{\text{true}})$ can be allocated on a fixed position as sketched in Figure 2. In the $\mathcal{M}$-open case, this is formally not possible, but since the framework of Pseudo-BMS/BMA allows to handle conceptual uncertainty for predictive purposes also for evolving model ensembles, it is considered to be suitable also in $\mathcal{M}$-open (Yao et al., 2017; Akaike, 1973). However, even if Pseudo-BMS/BMA were applied in an $\mathcal{M}$-closed setting, it is not able to identify a true model since it is tailored for other $\mathcal{M}$ settings. It might select the true model from the ensemble as best predictive one, but it does not allow for concluding that it is also the true model because it lacks the consistency property.

Regardless of the $\mathcal{M}$-setting in which Pseudo-BMS/BMA is employed, it is crucial to keep its information-theoretic interpretation in mind. The posterior marginalized likelihood is a particular instance of posterior predictive density $p(\boldsymbol{D}'|\boldsymbol{D}, M_m)$ and rates conditioned (calibrated) but wrong models in their ability to approximate the observations. It can be interpreted as so-called posterior predictive check (Gabry et al., 2017) at $\boldsymbol{D}'$ and is the suitable metric to rate models with a predictive purpose.

2.3.3 Bayesian Stacking

Model Combination: While the selection of a single model is the logical goal in the $\mathcal{M}$-closed setting where the true model can be identified, it might not generally be the best option for handling multiple models outside of it - because there, even the best predictive model is wrong. Therefore, alternatively to selecting the best single (posterior) predictive model as in Pseudo-BMS/BMA, it might be better to use several models together and declare the combination of model pdfs as final goal. Ideally, they complement one another and their combination achieves higher predictive capability than either single candidate. It might be plausible that the true model is "encircled" by the models in the ensemble. We can imagine this as a convex hull of model output distributions that contains the true distribution $q(\boldsymbol{y}|M_{\text{true}})$ (Sanderson et al., 2015) (see Figure 2) in the $\mathcal{M}$-complete setting. Then, a multi-model framework is required that yields model weights such that the weighted average rather than a single model matches the "encircled" truth.

A probabilistic framework that superposes or "stacks" predictive distributions of all models in the ensemble and optimizes the model weights according to how the whole combination performs is *Bayesian Stacking*. It optimizes the model weights

according to how well the combined or stacked (posterior) predictive distribution matches $q(\boldsymbol{y}|M_{\text{true}})$. This is opposed to Pseudo-BMS/BMA (Section 2.3.2) and BMS/BMA (Section 2.3.1) that rate the (posterior or prior) predictive distribution of each single model in terms of how well it matches $q(\boldsymbol{y}|M_{\text{true}})$.

Being outside of $\mathcal{M}$-closed, the same pointwise predictive densities as in Pseudo-BMA/BMS are used in Bayesian Stacking. The optimal weights are found by maximizing the combined predictive density under the log-score following:

$$\hat{\boldsymbol{w}}_{\text{stack}} = \arg\max_{\boldsymbol{w}} \frac{1}{N_s} \sum_{o=1}^{N_s} \ln \sum_{m=1}^{N_M} w_m p(D_o|\boldsymbol{D}_{\varnothing}, M_m) \tag{20}$$

Equation 20 is subject to $w_m \geq 0$ and $\sum_{m=1}^{N_M} w_m = 1$ (convex a.k.a. simplex constraint) and therefore resembles a simple linearly constrained maximization problem that can be solved by using a Lagrange multiplier. The combination of N_M model candidates writes as convex set (Yao et al., 2017):

$$\mathcal{K} = \left\{ \sum_{m=1}^{N_M} w_m p(\boldsymbol{y}|\boldsymbol{D}, M_m) | \sum_{m=1}^{N_M} w_m = 1, w_m \geq 0 \right\} \tag{21}$$

The model weights in $\mathcal{K}$ are no model probabilities despite their simplex constraint. In Bayesian Stacking, they represent the optimal shares $\hat{\boldsymbol{w}}_{\text{stack}}$ of each model in a convex combination under the chosen scoring rule. In case another score than the D_{KL} is chosen for optimizing model weights (e.g. the Brier score), also another objective function than in Equation 20 has to be maximized and the optimal weights differ (Yao et al., 2017).

Originally, Stacking (Wolpert, 1992; Breiman, 1996) is a general concept for obtaining a weighted average of "best" (point) estimates $\hat{\boldsymbol{y}}_m$ from multiple calibrated models M_m. The weights are typically gained by minimizing the error metric SSE between $\hat{\boldsymbol{y}}_m$ and observations $\boldsymbol{D}$. Then the best point estimates $\hat{\boldsymbol{y}'}_m$ from all models are accordingly weighted to match new data $\boldsymbol{D}'$. For such point estimator combination, the simplex constraint on the weights can be relaxed and negative weights can be used (Bates and Granger, 1969) while in Bayesian stacking the constraint is a natural choice (Yao et al., 2017).

Model Averaging: The obtained stacking weights provide an average (weighted mixture) of posterior predictive pdfs in the actual sense of pdf combination. Yet, not single model predictions are averaged as in classic stacking, but their distributions. This means that in the prediction space $\mathcal{Y}$, the individual model pdfs do not change their position but their weighted fractions form a hull to cover the

unknown $q(\boldsymbol{y}|M_{\text{true}})$.

The model weights do not express conceptual uncertainty in terms of relative preference for one model over the others. Conceptual uncertainty is accounted for by admitting each model candidate a certain share in the combination, but it is not meant to be "resolved" under growing data size as in BMS/BMA and Pseudo-BMS/BMA. For the $\mathcal{M}$-complete setting, the model weights are supposed to converge to static model weights such that, optimally, $q(\boldsymbol{y}|M_{\text{true}})$ lies within the convex hull created by the models in the ensemble. In the $\mathcal{M}$-open case, since $q(\boldsymbol{y}|M_{\text{true}})$ cannot be conceptually allocated, Bayesian Stacking might yield better predictive coverage by employing combined models for a certain amount of data but under growing data size the model weights might change as well.

Interpretation: In Bayesian Stacking, the weighted average of models is closest to the unknown truth in an information-theoretic sense, yet not necessarily matching it. Being interpreted from the perspective of scores and divergences (cf. Section 2.2.4), Bayesian stacking yields model weights to minimize the D_{KL} between their mixed posterior predictive pdf $p(\boldsymbol{y}|\boldsymbol{D})$ and the unknown $q(\boldsymbol{y}|M_{\text{true}})$ (Yao et al., 2017). Le et al. (2017) showed that $\mathcal{K}$ minimizes the information-theoretic loss between these two distributions under the log-score. Thus, it is the optimal Bayesian estimator outside of an $\mathcal{M}$-closed setting.

2.4 Implementation of Bayesian Inference

2.4.1 Inferring Distributions and Normalizing Constants

Bayesian inference for mathematical models requires to integrate out the parameter distribution in order to obtain posterior predictive distributions and marginalized likelihoods. Sometimes, this can be conducted analytically when a certain likelihood is used with a so-called conjugate prior (Schlaifer and Raiffa, 1961). Then, a corresponding posterior can be analytically inferred (Box and Tiao, 1973). This holds for the variables of individual models and, in $\mathcal{M}$-closed, even on both levels of inference (MacKay, 1992), i.e., also for the distribution over all models. This conjugacy holds mostly for distributions of the exponential family like Binomial or Gaussian, as respective examples for discrete and continuous distributions. The Gaussian distribution is even self-conjugate: A Gaussian prior is a conjugate prior for the Gaussian likelihood (Equation 8), and together they automatically yield a Gaussian posterior as analytical solution to Bayesian inference. A small collection of other conjugate examples (e.g., DeGroot, 2005) is given in Table 2.

Table 2: Examples of conjugate distributions: corresponding posterior distributions are analytically tractable if the specified likelihood is used with a conjugate prior.

likelihood $p(\boldsymbol{D}\|\boldsymbol{\Theta}_m)$	conjugate prior $p(\boldsymbol{\Theta}_m)$	posterior $p(\boldsymbol{\Theta}_m\|\boldsymbol{D})$
Gaussian	Gaussian	Gaussian
Poisson	Gamma	Gamma
Binomial	Beta	Beta
Categorical	Dirichlet	Dirichlet
Multinomial	Dirichlet	Dirichlet

However, such analytical solutions are rarely applicable. A typical example is a Gaussian likelihood in linear regression with normally distributed regression parameters. Usually, as for nonlinear models, Bayesian inference becomes analytically intractable and other methods have to be employed.

Most of these methods are based on statistical sampling of the distributions. In principle, the methods exploit that, first, Bayesian inference and uncertainty quantification rests upon the proportionality: *posterior* $\propto$ *likelihood* $\cdot$ *prior* (see Equation 4); and, second, that Bayesian model rating rests upon turning this proportionality into equality by marginalization.

The most straight-forward numerical approach that provides both is plain Monte Carlo (MC) integration (e.g., Hammersley and Handscomb, 1964). With MC we numerically draw N_{MC} independent samples ζ_i from a random variable Z with distribution of interest $q(\zeta)$, e.g., a prior parameter pdf. With these samples, we can approximate the expected value $\mathrm{E}_q[f(\zeta)]$ of a function $f(\zeta)$ (e.g., a likelihood function) over $q(\zeta)$ that is defined as:

$$\mathrm{E}_q[f(\zeta)] = \int f(\zeta)q(\zeta)d\zeta \approx \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} f(\zeta_i) \tag{22}$$

BME is the likelihood function $p(\boldsymbol{D}|\boldsymbol{\Theta}_m)$ integrated (marginalized) over the parameter prior pdf $p(\boldsymbol{\Theta}_m)$ and can therefore be evaluated with plain MC via Equation 22. MC is computationally demanding but reliable in converging to the expectation. Convergence of the mean to the expectation is guaranteed by the law of large numbers (see Section 3.2.2). Therefore, I employ MC in this thesis (see Chapter 4).

Advanced (and related) methods like Markov Chain Monte Carlo (MCMC) allow for sampling of $q(\zeta)$ if it is not directly tractable (see, e.g., Andrieu et al., 2003) but can be evaluated at ζ_i. Briefly, MCMC "jumps" over the target distribution and stores thereby accepted samples in a so-called chain. The acceptance of samples follows strict rules that assure convergence to the target distribution. In Bayesian

inference, these jumps are in accordance with the above proportionality. Hence, the posterior is sampled and resembled by the collection of samples in the chain - yet, only up to an unknown constant. For normalization, further processing of the samples has to be applied (e.g. Vehtari et al., 2017). Such advanced techniques enable us to efficiently perform Bayesian inference by accessing the involved distributions by more sophisticated sampling strategies than plain MC, e.g., some MCMC techniques employ several chains. They usually provide more informative samples faster but, as trade-off, they typically have methodical parameters that require fine-tuning. Despite not being employed here, I provide more specifics on numerical methods and their ties to Bayesian inference in Appendix A.

2.4.2 Bayesian Bootstrap

Like every other numerically estimated quantity, the model weights from BMS/BMA, Pseudo-BMS/BMA and Bayesian Stacking are subject to inferential uncertainty:

- First, the evaluation of weights in either method rests on quantities that are marginalized over the whole considered parameter distribution ($p(\boldsymbol{\Theta}_m)$ or $p(\boldsymbol{\Theta}_m|\boldsymbol{D}_\varnothing)$). With these distributions being approximated by numeric samples, there is always uncertainty about the convergence of the marginalized quantities. This is a question of appropriate sampling algorithms and sufficient numerical sample sizes to assure full convergence (Schöniger et al., 2014).

- Second, it remains unclear whether the used observations $\boldsymbol{D}$ are a sufficient proxy for the whole unknown data distribution $q(\boldsymbol{y}|M_{\text{true}})$ (see the problem of finite data (Nearing and Gupta, 2018) in Section 1.1). Especially in predictive model selection or combination like Pseudo-BMS/BMA and Stacking, this uncertainty propagates to the estimated model weights and has to be accounted for, e.g., via a Bayesian Bootstrap approach as in Yao et al. (2017) for statistical models.

The Bayesian Bootstrap (BB) introduced by Rubin (1981) can be considered as Bayesian analogue of the Frequentist bootstrapping (Efron, 1979). It evaluates the uncertainty of sampled distributions by resampling (Efron and Tibshirani, 1994). Both provide a non-parametric approximation to the distribution of a random variable. The BB employs a uniform Dirichlet distribution, i.e., a distribution over distributions: The data-points themselves stem from the data distribution $q(\boldsymbol{y}|M_{\text{true}})$. Every available data-point D_o in $\boldsymbol{D}$ is a sample from this distribution. Yet, $\boldsymbol{D}$ resembles only one instance of $\boldsymbol{y}$ that follows $q(\boldsymbol{y}|M_{\text{true}})$. Hence, also any derived quantities like $\ln p(D_o|\boldsymbol{D}_\varnothing, M_m)$ in Pseudo-BMS/BMA are only a special case. In order to randomize this instance and subsequently infer statistics on the

randomized samples, the Dirichlet distribution lends itself to be suitable.

The Dirichlet distribution is the conjugate prior (cf. Section 2.4.1) to the posterior distribution for both, the multinomial and the categorical (a special case of the multinomial distribution) likelihood distribution (see Table 2). Let us take i.i.d. elements in a vector $\boldsymbol{\zeta}$ as samples of some random variable Z. Via the Dirichlet distribution, their occurrences are assigned a posterior probability of 1 since they are contained in $\boldsymbol{\zeta}$, expressed as $Dirichlet(\mathbf{1})$ with $\mathbf{1} \equiv (1, ..., 1)$ of length N_s. Samples of Z that are not in $\boldsymbol{\zeta}$ are assigned 0 probability since they have no probability under the sample cumulative distribution function (Rubin, 1981). Figuratively spoken, each sample has its own bin. The prior distribution of samples taken from these bins resembles a categorical distribution - one "category" for each bin. The multinomial distribution generalizes this over multiple drawings from all these bins; and the Dirichlet distribution depicts the respective posterior.

The BB is explained in the following for $\boldsymbol{\zeta}$ being the vector of logarithmic LOO predictive densities of model M_m, i.e., with $\zeta_{o,m} = \ln p(D_o|\boldsymbol{D}_{\varnothing}, M_m)$. Per Bootstrapping replication b with $b = 1 : N_{BB}$ the drawn posterior probabilities $\boldsymbol{\alpha}_{1:N_s,b}$ for $\boldsymbol{\zeta}$ follow (Yao et al., 2017):

$$\boldsymbol{\alpha}_{1:N_s,b} \sim Dirichlet(\mathbf{1}) \tag{23}$$

In the Bayesian Bootstrap procedure, we now draw N_{BB} statistically plausible and varying alternatives of $\boldsymbol{\zeta}$. These so-called Bootstrapping replications yield the sampling-based Bootstrapping distribution of the distribution of Z over which any statistical moment can be inferred (Rubin, 1981) - for instance the mean (Yao et al., 2017):

$$\bar{\zeta}_{b,m} = \sum_{o=1}^{N_s} \alpha_{o,b}\zeta_{o,m} \tag{24}$$

Thereof, the N_{BB} replicates of the model weight of model M_m are simply estimated by:

$$w_{b,m} = \frac{\exp(N_s\bar{\zeta}_{b,m})}{\sum_{m=1}^{N_M}\exp(N_s\bar{\zeta}_{b,m})} \tag{25}$$

The expected weight over the whole BB distribution then writes as:

$$w_m^{\mathrm{BB}} = \frac{1}{N_{BB}}\sum_{b=1} w_{b,m} \tag{26}$$

Bootstrapping typically counteracts extreme weights of 0 or 1 (Yao et al., 2017). The major strength of the BB is, however, that it allows to formulate likelihood statements about the moments of the BB distribution (Rubin, 1981). This means that the BB mean of weights w_m^{BB} is - accounting for the uncertainty in definiteness of data $\boldsymbol{D}$ - more likely than the direct calculation of weights without bootstrapping. In case the weights calculated without bootstrapping are the same as after applying it, bootstrapping can be seen as confirmation. The additional computational costs of the BB are very small because the quantities required to apply it (here, $\zeta_{o,m} = \ln p(D_o|\boldsymbol{D}_\varnothing, M_m)$) are already available for calculating the model rating scores.

2.5 Approximative Model Rating Methods

The marginalized likelihoods $p(\boldsymbol{D}|M_m)$ and $p(\boldsymbol{D}'|\boldsymbol{D}, M_m)$ (or more specifically $p(D_o|\boldsymbol{D}_\varnothing, M_m)$) from Section 2.3 each follow the law of parsimony (see Section 1.1): They account for model complexity in their respective realm while quantifying the overall model fit to data $\boldsymbol{D}$ (or $\boldsymbol{D}'$). Note, that when these two quantities are obtained by full marginalization, e.g., via MC integration, model complexity is taken into account implicitly.

Approximative estimators of the logarithmic marginalized likelihoods are commonly used and known as model selection criteria, most of them are called information criteria (IC) (e.g. Spiegelhalter et al., 2014). They resemble decompositions of the logarithmic scores into a term for so-called goodness-of-fit to $\boldsymbol{D}$ and a specific model complexity part. Hence, they depict how model complexity is quantified with each score explicitly.

In the following, commonly used Bayesian model selection criteria (and their non-Bayesian counterparts) are presented and discussed. Thereby, the model-specification M_m is dropped for simplification. For clarity, the criteria are assigned to four classes according to Höge et al. (2018). These classes distinguish them by whether they are consistent (B-type) or non-consistent (A-type) and by whether they are Bayesian or non-Bayesian (1 or 0). Criteria in class:

- B1 approximate $\ln p(\boldsymbol{D})$. Models are rated by posterior model probability (class of BMS/BMA, see Section 2.3.1).
- B0 resemble the non-Bayesian counterpart to B1. Models are rated by code length that is approximated by respective criteria.
- A1 approximate $\ln p(\boldsymbol{D}'|\boldsymbol{D})$. Models are rated by predictive density (class of Pseudo-BMS/BMA, see Section 2.3.2).

- A0 resemble the non-Bayesian counterpart to A1. Models are rated by predictive error that is approximated by respective criteria.

The rest of the remaining Chapter 2 has been published by Höge et al. (2018) and I reuse parts of the text.

In their original derivation, many model selection criteria assume that residuals between observations $\boldsymbol{D}$ and model predictions $\boldsymbol{y}$ can be described as white noise (zero mean, uncorrelated, finite variance) or even as independent and identically distributed (i.i.d.) (e.g. Leeb and Pötscher, 2009). This holds for the uncorrelated Gaussian case but rarely occurs in reality - especially in hydro(geo)logy. Hence, a reasonable treatment of the errors is required when model selection criteria are applied (e.g. Schoups and Vrugt, 2010). However, it was shown that model selection criteria generally work under weaker assumptions on the errors than being Gaussian or i.i.d. (Leeb and Pötscher, 2009, and references therein). Principally, it has to be noted that all criteria are conditional on the choice of the error distribution (also known as loss, cost or likelihood function) (e.g. Tarantola, 2006).

2.5.1 B1: Posterior Model Probability

For linear models with Gaussian parameter prior and Gaussian likelihood, the Kashyap information criterion KIC (Kashyap, 1982) provides an analytically correct solution to the marginal likelihood given by $p(\boldsymbol{D}) = \exp(-\frac{1}{2}\mathrm{KIC})$. It is based on the Laplace approximation about the maximum a posterior estimator (MAP) $\tilde{\boldsymbol{\Theta}}$ (Schöniger et al., 2014) and it can be applied whenever this approximation is valid:

$$\mathrm{KIC}_{\mathrm{MAP}} = -2\ln\mathcal{L}(\tilde{\boldsymbol{\Theta}}|\boldsymbol{D}) \underbrace{-2\ln p(\tilde{\boldsymbol{\Theta}}) - N_p \ln(2\pi) - \ln|\tilde{\boldsymbol{\Sigma}}|}_{\mathrm{lnOF}_{\mathrm{KIC}}} \tag{27}$$

The KIC allows for splitting up the logarithmic marginal likelihood explicitly into a goodness-of-fit term (first term in equation 27) and a so-called logarithmic Occam factor (OF) (MacKay, 1992) comprising three complexity terms. A detailed discussion on the effect of each of these terms can be found in Schöniger et al. (2014). In summary, the first two of these terms penalize complexity with respect to the number and prior uncertainty of parameters and balance each other partially by mutual compensation. The last term can be interpreted as a penalty for low parameter sensitivity towards data, i.e. for poor parameter identifiability by the given data $\boldsymbol{D}$.

As an alternative to the KIC evaluated at the MAP, KIC is frequently evaluated at the maximum likelihood estimator (MLE) (Neuman, 2003; Ye et al., 2004;

Schöniger et al., 2014). For larger sample sizes it is assumed that the likelihood function dominates the posterior parameter distribution and the MAP approaches the MLE. Hence, for these cases, BME can be reasonably approximated by evaluating the KIC terms at the MLE for large (informative) data sets.

The popular Schwarz or Bayesian IC (SIC/BIC; Schwarz, 1978) is the most compact approximation to BME. The BIC is derived in the limit of infinite sample size $N_s \to \infty$. Then again, the MLE $\hat{\boldsymbol{\Theta}}$ becomes equal to the MAP. In this limit, all Occam factor terms in equation 27 that are not affected by N_s drop because they become negligible. The Occam factor that remains is $N_{p,k}\ln(N_s)$. The BIC therefore writes as:

$$\mathrm{BIC} = -2\ln\mathcal{L}(\hat{\boldsymbol{\Theta}}|\boldsymbol{D}) \underbrace{+N_p\ln(N_s)}_{\ln\mathrm{OF}_{\mathrm{BIC}}} \tag{28}$$

In theory, the BIC converges to BME in the limit of infinite sample size. In practice, it is criticized for yielding unsatisfactory results even for large N_s (Kass and Raftery, 1995). Nonetheless, the BIC is the most popular consistent information criterion due to its simplicity. Hence, the whole branch of consistent model selection is often referred to as BIC-type model selection (Aho et al., 2014). Like the AICc in non-consistent model selection, there is a proposed correction for small sample sizes called BICc (McQuarrie, 1999).

So far, reliable BME evaluation metrics do either underlie strong assumptions like the above explicit criteria, or they are computationally demanding like the mentioned implicit schemes. Therefore, it is not possible to measure model complexity in the above sense explicitly when these assumptions do not hold (Schöniger et al., 2014). Criteria like the Watanabe-Bayesian IC (WBIC) (Watanabe, 2013) have been proposed to resolve this issue, but in various cases they perform poorly in approximating the BME when tested against implicit methods that directly assess BME (see e.g. Mononen, 2015; Friel et al., 2016).

In summary, model complexity quantified as Occam factor OF as part of the BME is the knowledge gain between parameter prior and posterior. It grows with the data size N_s and only parameters that are affected by $\boldsymbol{D}$ (i.e. are identified by $\boldsymbol{D}$) contribute to this value.

2.5.2 B0: Code Length

In coding theory, a model is considered to be "a compact representation of possible data one could observe" (Ghahramani, 2013). The coding-theoretic Kolmogorov(-Chaitin) complexity (KC) formalizes this concept of a model by evaluating the

complexity of a sequence (Grünwald, 2000). KC is the shortest code in bits that can produce a certain output, e.g. a sequence of symbols like a series of data, and then halts (Grünwald and Vitányi, 2003). For reasons not further discussed here, KC is considered to be incomputable (Rathmanner and Hutter, 2011).

From a coding theory point of view, everything can be encoded. In this spirit, fitting a model can be considered as encoding the data. The shortest coded compression of data $\boldsymbol{D}$ is the simplest statistical model that can reproduce $\boldsymbol{D}$ (Rissanen, 1978; Grünwald, 2000). The idea of compressing data is based on the assumption that there is pattern or structure in the observations. A set of data without any structure cannot be compressed easily and each data point has to be stored explicitly. This enlarges the code and makes the required model more complex. The more compression due to redundancy or structure is possible, the better a simple model can describe the regularities behind the observations (Myung et al., 2000).

This perspective motivated the development of model selection based on Minimum Description Length (MDL) (Rissanen, 1978). The MDL of a model candidate to compress $\boldsymbol{D}$ is its code length needed (Myung et al., 2006). The popular version of MDL presented here is formalized as the so-called Fisher information approximation (e.g. Vandekerckhove et al., 2015) (see section 2.5.3 for details on the Fisher information matrix $\mathbf{F}$):

$$\text{MDL} = -\ln\mathcal{L}(\hat{\boldsymbol{\Theta}}|\boldsymbol{D}) \underbrace{+\frac{N_p}{2}\ln\frac{N_s}{2\pi} + \frac{1}{2}\ln\int|\mathbf{F}_1(\boldsymbol{\Theta})|\mathrm{d}\boldsymbol{\Theta}}_{\text{GC}} \tag{29}$$

The MDL consists of two parts (e.g. Barron et al., 1998; Myung et al., 2000): A first part represents the code length that is needed to describe the deviations between data and best-fit model predictions (goodness-of-fit). A second part encodes the functional relations of the model and its parameters, i.e. the complexity of the model, called geometric complexity (GC). The idea behind GC is that a model generates (likelihood) distributions $p(\cdot|\boldsymbol{\Theta})$. A model therefore represents a "family of probability distributions consisting of all the different distributions that can be reproduced by varying $\boldsymbol{\Theta}$" (Myung et al., 2006). The complexity of the model then refers to how similar these distributions are (Rissanen, 1996). A model is considered to be simple, if the distributions are hardly distinguishable: less distinguishability means more structure in the data and more structure means more compressibility in code (Myung et al., 2000).

In so-called entropy encoding, code length is approximately proportional to the negative logarithmic probability density (Friedman et al., 2001; Myung et al., 2006). A high density means little deviations or small errors, which in turn require just short pieces of code to be compressed (Barron et al., 1998). Hence, the goodness-of-fit term and the GC terms in equation 29 can be interpreted as code lengths.

GC in equation 29 can be seen as the logarithm of the counted number of distinguishable distributions over the model's whole parameter space, hence growing with N_p. The counting is based on a differential-geometric distance measure which employs the Fisher information matrix normalized by the number of observations $\mathbf{F}_1 = \mathbf{F}\, N_s^{-1}$. With this metric it is possible to quantify how "close" the distributions are, i.e. whether they can be distinguished and counted separately or not - for more details refer to e.g. Myung et al. (2000) and references therein.

In summary, GC is a coding-theoretic counter for distinguishable distributions produced by the model and it grows with data size N_s. MDL selects the model with the highest ratio between goodness-of-fit and the number of distributions the model can generate (Myung et al., 2000). It can be used for non-Bayesian consistent model selection (Lanterman, 2001) but is numerically demanding in case no closed-form solutions for the evaluation of GC are available.

2.5.3 A1: Predictive Density

Model selection criteria which try to estimate predictive density assume that there is a (infinite dimensional) true model with a predictive density function $q(\boldsymbol{y})$ for an observable random variable $\boldsymbol{y}$. They are called predictive information criteria (IC). The exact pdf $q(\boldsymbol{y})$ is generally unknown, but the observed (within-sample) data $\boldsymbol{D}$ and future (out-of-sample) data $\boldsymbol{D'}$ are both assumed to follow $q(\boldsymbol{y})$.

However, without actual out-of-sample data $\boldsymbol{D'}$, predictive IC can only approximate $\mathrm{E}_q[\ln p(\boldsymbol{D'}|\boldsymbol{D})]$ using the given within-sample data $\boldsymbol{D}$. This results in an offset (Hooten and Hobbs, 2015) that is caused by testing a model on the data set on which it was conditioned (fitted). Predictive IC incorporate this offset by an effective number of parameters N_p^* and use it as complexity representation of the model (Akaike, 1973). This correction can be interpreted as a quantification of how much predictive density for $\boldsymbol{D'}$ increases by fitting N_p^* parameters to $\boldsymbol{D}$ (Gelman et al., 2014).

The most popular predictive IC is the Akaike information criterion AIC (Akaike, 1973, 1974, 1978). It is an approximation to the information-theoretic Kullback-Leibler-Divergence $\mathrm{D}_{\mathrm{KL}}(q, p)$ that quantifies the information loss between the pre-

dictive distributions of a hypothetical true model $q(\boldsymbol{y})$ and the candidate model $p(\boldsymbol{y})$ (Aho et al., 2014) in the model space:

$$\mathrm{D_{KL}}(q,p) = \int q(\boldsymbol{y}) \ln\left(\frac{q(\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{\Theta})}\right) \mathrm{d}\boldsymbol{y} = \mathrm{E}_q[\ln q(\boldsymbol{y})] - \mathrm{E}_q[\ln p(\boldsymbol{y}|\boldsymbol{\Theta})] \tag{30}$$

The first term on the right-hand side of equation 30 is a constant for all compared models. Therefore, the AIC addresses only the second term, called the relative expected KL-information (Burnham and Anderson, 2004), resembling the expected logarithmic predictive density. The approximation of equation 30 by the AIC was derived for asymptotic normal posterior distributions in the large-sample limit (e.g. linear models with uninformative parameter prior and normal error distribution). In this special case, the point-estimate $\hat{\boldsymbol{\Theta}}$ summarizes the posterior parameter distribution. Therefore, the model's expected log. predictive density is conditional on $\hat{\boldsymbol{\Theta}}$ and given by $\mathrm{E}_q[\ln p(\boldsymbol{D}'|\hat{\boldsymbol{\Theta}})]$ (Gelman et al., 2014). This cannot be directly calculated, but with all candidate models being conditioned on the same data $\boldsymbol{D}$, it can be approximated via the log.-likelihood value $\ln \mathcal{L}$ evaluated at the model-specific MLE $\hat{\boldsymbol{\Theta}}(\boldsymbol{D})$ plus a correction for the approximation offset. Under the above conditions, this correction naturally appears in the derivation as simply the number of model parameters N_p (e.g. Burnham and Anderson, 2002). Hence, the AIC writes as (with a factor of two for historical reasons):

$$\mathrm{AIC} = -2\ln\mathcal{L}(\hat{\boldsymbol{\Theta}}|\boldsymbol{D}) + 2\,N_p \tag{31}$$

A model with many parameters can reduce $\mathrm{D_{KL}}(q,p)$ for $\boldsymbol{D}$ by fitting all of these N_p parameters. Since the AIC was derived for uninformative priors, all the model parameters have to be constrained by $\boldsymbol{D}$ instead of prior information. Hence, in equation 31, the goodness-of-fit (negative first term) in the model selection criterion has to be reduced by this "potential" for reducing $\mathrm{D_{KL}}(q,p)$ (positive second term), i.e. the independently adjustable parameters (Akaike, 1974). Interestingly, the factor of two converts the first term in equation 31 to a plain sum of squared errors for an uncorrelated normal error distribution. A version of the AIC corrected for finite data size N_s (AICc) was developed (Hurvich and Tsai, 1989) to compensate for smaller sample sizes in case of which the above asymptotic behaviour cannot be assumed, i.e. N_s is too small that the IC could reliably select the model with the largest predictive capability.

A generalization of the AIC was proposed as Deviance information criterion (DIC) (Spiegelhalter et al., 2002). In contrast to the AIC, the DIC was designed for informative priors and can therefore be seen as a more Bayesian version of the AIC (Spiegelhalter et al., 2014). The deviance $D(\boldsymbol{\Theta})$ is defined as the doubled

negative logarithmic likelihood (NLL): $D(\boldsymbol{\Theta}) = -2\ln\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{D})$ (as it was used for the AIC). The DIC is evaluated at the posterior parameter mean $\overline{\boldsymbol{\Theta}}$:

$$\text{DIC} = -2\ln\mathcal{L}(\overline{\boldsymbol{\Theta}}|\boldsymbol{D}) + 2\,N_p^* \tag{32}$$

In contrast to AIC, model complexity is measured as *effective* number of parameters N_p^*, which does not necessarily equal the straightforward parameter count N_p. The DIC does not require asymptotic normality in the large-sample limit. This extends the applicability of the DIC to non-linear models and to incorporate informative priors (in contrast to AIC) as long as the posterior parameter distribution can be sufficiently approximated by a Gaussian even under limited sample size. Being evaluated at the posterior mean $\overline{\boldsymbol{\Theta}}$, the DIC uses an averaged quantity based on the assumed normality. In principle, this is more Bayesian than just using the MLE as point-estimate, but it relies heavily on the Gaussian assumption. This is not a real marginalization (see 2nd level of Bayesianism) and makes the DIC subject to criticism (e.g. Piironen and Vehtari, 2017).

The derivation of N_p^* in equation 32 based on the deviance $D(\boldsymbol{\Theta})$ is as follows: If the posterior parameter distribution is multivariate Gaussian, the deviance automatically follows a $\mathcal{X}^2$ distribution. This is typically given for errors being normally distributed (Clark and Gelfand, 2006). As a property of the $\mathcal{X}^2$ distribution, the difference between the mean density $\overline{D(\boldsymbol{\Theta})}$, and the density at the mean, $D(\overline{\boldsymbol{\Theta}})$, is equal to the statistical degrees of freedom ν of the $\mathcal{X}^2$ distribution. The DIC uses this difference to approximate ν and then defines it as the number of effective parameters N_p^* (Spiegelhalter et al., 2002):

$$N_{p,DIC1}^* \equiv \nu \approx \overline{D(\boldsymbol{\Theta})} - D(\overline{\boldsymbol{\Theta}}) \tag{33}$$

Exploiting another property of the $\mathcal{X}^2$ distribution, Gelman et al. (2004) suggested to use half of the variance of the Deviance over the posterior to estimate the effective number of parameters. This is possible, because just like the difference in equation 33, $\frac{1}{2}\text{Var}[D(\boldsymbol{\Theta})]$ also equals the distribution's statistical degrees of freedom:

$$N_{p,DIC2}^* \equiv \nu \approx \frac{1}{2}\text{Var}[D(\boldsymbol{\Theta})] \tag{34}$$

Spiegelhalter et al. (2002) describe $N_{p,DIC1}^*$ or $N_{p,DIC2}^*$ as the dimension of parameter space that can be constrained by the given data, calling it a model dimensionality. Since N_p^* is not necessarily equal to N_p, it shall not be confused with parameter space dimensionality. However, N_p^* reduces to N_p if the prior is uninformative (van der Linde, 2012; Meyer, 2014) and the DIC reduces to the AIC.

The AIC and the DIC are limited to so-called regular models. This means that certain regularity conditions hold, e.g. the Fisher information matrix $\mathbf{F}$ exists and is positive definite (Watanabe, 2010). Otherwise, a model is called singular. $\mathbf{F}(\boldsymbol{\Theta})$ is defined as the negative Hessian of the log-likelihood $\ln\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{D})$ with respect to parameters: $\mathbf{F}(\boldsymbol{\Theta}) = -\frac{\partial^2 \ln\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{D})}{\partial\boldsymbol{\Theta}^2}$. The inverse of $\mathbf{F}(\boldsymbol{\Theta})$ is an estimator to the posterior covariance matrix among the parameters. For singular models, $\mathbf{F}$ is not positive definite. Hence, there can be parameters with infinite variance even after calibration.

Because the AIC and DIC are limited to regular models (van der Linde, 2012), the "widely-applicable" (or "Watanabe-Akaike") information criterion (WAIC) was developed (Watanabe, 2010) as generalization of the AIC and DIC to singular models (Betancourt, 2015) such as Gaussian mixture models, strongly over-parametrized models (causing under-determined inverse problems) (Gelman et al., 2004) or artificial neural networks (Watanabe, 2010). The WAIC writes as:

$$\mathrm{WAIC} = -2\sum_{i=1}^{N_s} \ln(\mathrm{E}[p(D_i|\boldsymbol{\Theta})]) + 2\,N_p^* \tag{35}$$

Again, model complexity is measured by the effective number of parameters in two versions, called $N^*_{p,WAIC1}$ and $N^*_{p,WAIC2}$ in the following. $N^*_{p,WAIC1}$ is estimated in a similar way as $N^*_{p,DIC1}$ in equation 33. However, for $N^*_{p,WAIC1}$ the difference is evaluated for each observation D_i in $\boldsymbol{D}$ independently over the whole parameter space, and then summed over all N_s observations, approximating leave-one-out (LOO) cross-validation (CV) - details on LOO follow at the end of this section).

$$N^*_{p,WAIC1} = 2\sum_{i=1}^{N_s} (\ln(\mathrm{E}[p(D_i|\boldsymbol{\Theta})]) - \mathrm{E}[\ln p(D_i|\boldsymbol{\Theta})]) \tag{36}$$

Similarly to $N^*_{p,DIC2}$ above, a variance-based estimator of the N_p^* exists, yielding a second version of the WAIC:

$$N^*_{p,WAIC2} = \sum_{i=1}^{N_s} \mathrm{Var}[\ln p(D_i|\boldsymbol{\Theta})] \tag{37}$$

It is still argued about whether the variance-based estimators in the DIC and the WAIC can be seen as generalizations of each other (Watanabe, 2010). However, for practical purposes, both are sometimes advantageous over the two difference-based estimators ($N^*_{p,DIC1}$ and $N^*_{p,WAIC1}$) because they cannot become negative (Gelman et al., 2014).

In the WAIC-related equations 35 to 37, expectations and variance of log predictive density are evaluated for each data point in $\boldsymbol{D}$ and then summed up. This is different from the approaches in the AIC and DIC where the log-likelihood function $\ln\mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{D})$ of the entire data set $\boldsymbol{D}$ is used. Further, with the underlying assumptions in the AIC and the DIC, they may use point-estimators to estimate the predictive density. The WAIC only uses quantities averaged over the whole parameter space and all independent observations. Therefore, the WAIC is considered the only fully Bayesian one among the predictive IC (Gelman et al., 2014).

In summary, model complexity quantified as effective number of parameters N_p^* is an offset correction for estimating the predictive density of unknown out-of-sample data $\boldsymbol{D}'$ by only using known within-sample data $\boldsymbol{D}$. N_p^* is conditional on $\boldsymbol{D}$ and can therefore be interpreted as the amount of parameters that are constrained by $\boldsymbol{D}$ rather than just by the prior information on parameters (Gelman et al., 2014).

2.5.4 A0: Predictive Error

An alternative starting point for assessing the predictive capability of a model is to interpret observables $\boldsymbol{y}$ deterministically rather than as random variables. Then, future data $\boldsymbol{D}'$ shall be met as exactly as possible and model predictions shall only show minimal residuals when compared to data. This is related but different from A1-type criteria that see data as samples of random variables with the goal for models to meet the right probability distributions.

In statistical learning (Friedman et al., 2001), calibration error and validation error are often referred to as training and test error, respectively. A common way to measure the training error is the residual sum of squares (RSS) $(\boldsymbol{D}-\hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{D}-\hat{\boldsymbol{y}})$ between observations $\boldsymbol{D}$ and the corresponding best-fit estimates $\hat{\boldsymbol{y}}$ of the model. Accordingly, the test error RSS writes as: $(\boldsymbol{D}'-\hat{\boldsymbol{y}}')^{\mathrm{T}}(\boldsymbol{D}'-\hat{\boldsymbol{y}}')$. As the potential future data $\boldsymbol{D}'$ are unknown, the test error has to be estimated by just using the training error from $\boldsymbol{D}$. As the training error underestimates the test error, a penalty term has to be added to the training error which accounts for the gap between the two types of errors. This term is assumed to be proportional to the so-called model degrees of freedom (DoF) (e.g. Friedman et al., 2001; Zou et al., 2007) - not to be confused with the statistical degrees of freedom ν in section 2.5.3. Adding these DoF, scaled by a known error variance σ_R^2, to the RSS yields a common estimate for test error, called the expected prediction error (EPE) (Janson et al., 2015):

$$\mathrm{EPE} = (\boldsymbol{D}-\hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{D}-\hat{\boldsymbol{y}}) + 2\sigma_R^2\,\mathrm{DoF} \tag{38}$$

This expression for EPE in equation 38 is also known as Mallows' C_p (Mallows,

1973; Efron, 1986; Janson et al., 2015). It can easily be seen that the RSS is proportional to a Gaussian log-likelihood, which leads to the same term (yet with a different derivation) as in the AIC. Therefore, the EPE is often interpreted similarly to the predictive information criteria from section 2.5.3 and model DoF are considered to be yet another measure for model complexity (Ye, 1998; Hooten and Hobbs, 2015). However, the DoF as model complexity are neither motivated by predictive density nor do they require any kind of Bayesian parameter prior.

DoF are meant to measure the sensitivity of model predictions $\boldsymbol{y}$ with respect to perturbations in the data $\boldsymbol{D}$ used for training (Ye, 1998). A model with high sensitivities does not allow for a unique calibration when calibrated to different data sets that even just slightly deviate from one another. This interpretation directly links to the stability of model inversion (e.g. Tarantola, 2005). In this sense, a model with high sensitivities to data is considered to have many degrees of freedom and is rated complex.

A widely accepted and used formulation for the DoF in model selection is the so-called expected optimism (Efron, 1983, 1986), assuming i.i.d. errors of finite variance σ_R^2:

$$\mathrm{DoF} = \frac{1}{\sigma_R^2} \sum_{i=1}^{N_s} \mathrm{cov}(\hat{\boldsymbol{y}}_i^*, \boldsymbol{D}_i^*) \tag{39}$$

In this approach, $\mathrm{cov}(\hat{\boldsymbol{y}}_i^*, \boldsymbol{D}_i^*)$ is estimated on repetitively perturbed data $\boldsymbol{D}^*$ and corresponding best-fit estimates $\hat{\boldsymbol{y}}^*$. This is a direct assessment of how sensitive model predictions are to noise in data. The perturbed data $\boldsymbol{D}^*$ can be obtained, for example, using residual bootstrapping (Efron and Tibshirani, 1994). DoF can generally be evaluated for linear and non-linear models (Janson et al., 2015). In the special case of Gaussian linear models, DoF is independent from data and predictions (Ye, 1998). Gaussian refers to the error distribution; and a linear model writes as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}$, with the parameter vector $\boldsymbol{\beta}$ and the independent variables being contained in the matrix $\mathbf{X}$ of base function vectors. The least-square estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\boldsymbol{D}$which yields $\hat{\boldsymbol{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. This can be reformulated by $\hat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{D}$, where the so-called projection matrix (a.k.a. hat, smoother or influence matrix) $\mathbf{S} = \mathbf{X}\,(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}$ describes the projection from observations to least-square estimates: $\hat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{D}$ (Cardinali et al., 2004). The diagonal elements of $\mathbf{S}$ are called leverages. The sum of leverages, i.e. the trace of $\mathbf{S}$, is interpreted as the model DoF. Thus, for linear models, equation 39 turns into $\mathrm{tr}(\mathbf{S})$ and yields the number of linearly independent predictors (Janson et al., 2015), i.e. the number of parameters: $\mathrm{DoF}_\mathrm{lin} = \mathrm{tr}(\mathbf{S}) = N_p$.

In general, the interpretation of model DoF has to be done carefully. In linear (polynomial) regression, DoF is equal to the number N_p of non-redundant free parameters in the model and are therefore accepted as model complexity measure (van der Linde, 2012). For nonlinear models, it is possible that the DoF are smaller or larger than the actual number of parameters (Janson et al., 2015). This counteracts our intuition of counting flexible parts of the model and makes it more important to consider model DoF as a complexity measure representing sensitivities rather than a parameter count. Following a similar spirit, methods exist (which can also be used for model selection) that bound the predictive error using, e.g., structural risk minimization (e.g. Friedman et al., 2001), the so-called covering number (Cucker and Smale, 2002) or related concepts (e.g. Pande et al., 2009, 2015).

In summary, the sensitivity-based DoF estimated from the available data $\boldsymbol{D}$ quantify the potential of poorly predicting new data $\boldsymbol{D}'$ due to unstable model inversion. This is a short-cut to classic (not Bayesian) cross-validation approaches (e.g. Friedman et al., 2001).

3 Investigating the Role of Model Complexity in Model Rating and Selection

The theory from Chapter 2 enables us to fully understand all parts of the central question from the introduction (Chapter 1): " Which multi-model framework employs the adequate Occam's razor with respect to the model setting of the modelling task at hand?". To begin with, we consider its core, i.e., how Occam's razor relates to the model setting of candidates in a set.

First, in Section 3.1, I show which practical relevance conceptual uncertainty has when a model is used to fit available old and potential new data and I discuss model complexity in this respect. In Section 3.2, I elucidate how and why model complexity is formalized as Occam's razor within the model selection criteria from Section 2.5 and its meaning in model rating. In Section 3.3, I introduce the developed general classification scheme that helps to find a suitable criterion in model selection. In Section 3.4, I provide a cross-comparison between the classes of model selection and an application example. I close the chapter with a summary and conclusion in Section 3.5.

3.1 Model Fit and Model Complexity

3.1.1 Overfitting and Underfitting

In statistical terms, conceptual uncertainty in modelling manifests itself as so-called overfitting or underfitting (e.g. Burnham and Anderson, 2002). An overfitted model closely matches available (within-sample) data $\boldsymbol{D}$, but struggles with reliably predicting (out-of-sample) data $\boldsymbol{D}'$. Such models are usually too flexible and tend to fit patterns in $\boldsymbol{D}$ that do not truly exist (e.g., are just caused by noise) which deteriorates the prediction of $\boldsymbol{D}'$. An underfitted model roughly meets the trend of $\boldsymbol{D}$, but struggles with following the actual pattern around the plain trend and therefore also in predicting $\boldsymbol{D}'$. Instead of being too flexible, an underfitted model is not flexible enough. As a rule of thumb, underfitting implies large bias, i.e., systematic error between model predictions and data, but typically also low variance of the predictions - overfitting implies the reverse. Bias and variance are visualized in Figure 5.

An illustration of both, overfitting and underfitting, is given by the simple binary classification problem in Figure 6, generated using *Python*'s *scikit-learn* package (Pedregosa et al., 2011). There, three different models shall classify 250 points (with two unspecified attributes on the axes) either by a red or blue label in a way that also further points will be correctly classified. We assume that the points

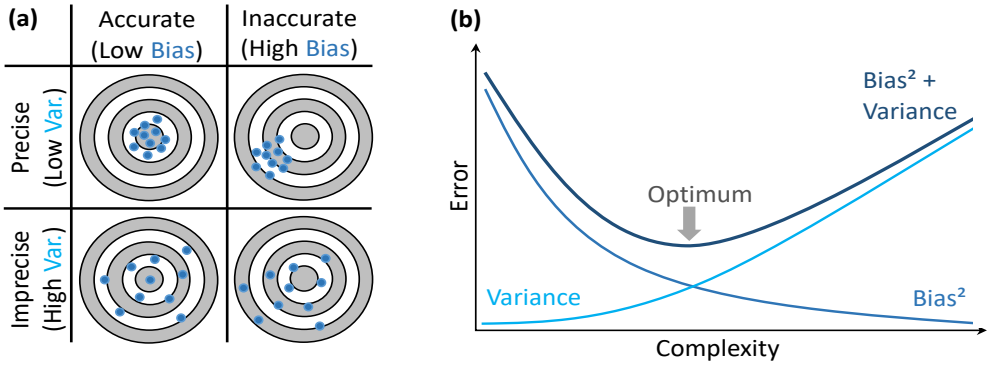


Figure 5: Concepts and effects of bias and variance: (a) Accuracy and precision visualized as bias and variance of shots on a target. Bias is the distance between the target center and the average position of shots. Variance is the spread of shots around their average. (b) Decomposition of total squared error into squared Bias and Variance (after e.g. Friedman et al. (2001)). Bias is supposed to decrease and Variance is supposed to grow with increasing model complexity, both due to growing model flexibility. Their superposition forms a minimum that marks optimal model complexity (from Höge et al., 2018).

are correctly separated by a smooth S-shaped curve with some noise that explains switched labels in the fringe zone. The model on the left separates the two classes by a straight line which clearly underfits the pattern of the data. The middle one appears as reliable estimation of the underlying classification model. The model on the right overfits and adopts also to noise rather than only the pattern of the data.

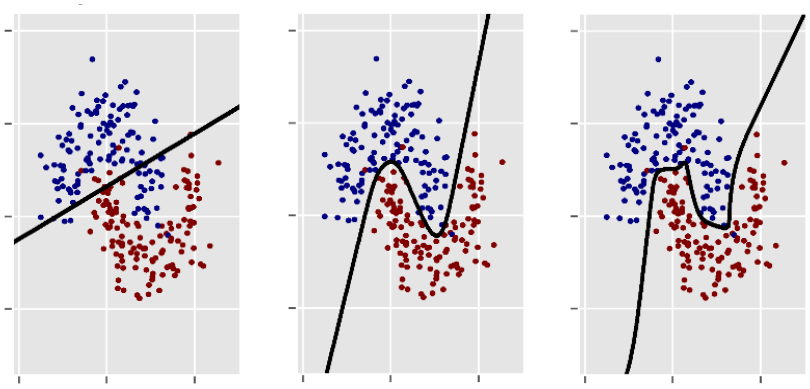


Figure 6: Illustrated underfitting (left), proper fitting (center) and overfitting (right) in binary classification of data to a red or blue class with unspecified attributes on the axes.

In regression, over- and underfitting can be easily illustrated by the following standard example: Being given 10 data points, a 9th order polynomial will yield perfect fit with zero residuals. Every lower-order polynomial will underfit and provide worse fit - the lower, the worse. Every higher-order polynomial will also perfectly fit the available data but also overfit: between the 10 points and when

leaving the within-sample data range, the "wiggling" of the polynomial model will become the stronger, the higher the order of the polynomial is. Note, that if the data are prone to error, the perfect fit of the 9th order polynomial is actually a misfit that can already be interpreted as overfit.

The fear to overfit or to underfit is typically implied when modellers refer to model complexity (see Figure 5). Typically, the excessive flexibility of overfitted models is assumed to come from too many parameters, functional terms, highly non-linear relationships, etc. They overestimate the complexity of the DGP and therefore fail to explain or to predict it. Overfitting poses a more frequently encountered problem than underfitting. It becomes apparent as, e.g., nonuniqueness of calibration or poor parameter identifiability (e.g. Schoups et al., 2008). Underfitting refers to the other extreme, where models underestimate the system complexity and are too simple to fully resolve the patterns of the DGP hidden in the data, i.e., to decipher the full system complexity.

3.1.2 Model Complexity Control

Reliable and successful modelling requires model complexity control (Schoups et al., 2008). I suggest to distinguish *within-model* and *between-model* complexity control:

- Within-model complexity control for a single model means limiting its flexibility.
- Between-model complexity control between multiple models (of typically deviating complexity) refers to either finding one model that suffers the least from overfitting or underfitting, or to employing models of different complexities together in order to mutually compensate individual shortcomings.

Within a model, complexity control is achieved by so-called regularization. This technique is applied throughout model calibration or conditioning in primarily ill-posed problems (e.g. Tarantola, 2005). Regularization means to provide further information to a model rather than only the data for calibration. Effectively, this additional information delimits the model output and therefore counteracts overfitting by reducing model flexibility or underfitting by preventing the extraction of false trends.

Typically, this additional information concerns the parameters and enables to constrain them during calibration, e.g., by preventing extreme parameter values. Common examples of regularization are the so-called LASSO or Tikhonov regularization (Marconato et al., 2013; Bardsley et al., 2015; Vaiter et al., 2015),

which respectively apply a L1- or L2-norm on the parameter values. When models are operated in a probabilistic (Bayesian) framework, they are assigned prior parameter distributions, which automatically act as such additional information. Therefore, applying a Bayesian prior is nothing else than putting a regularization on the model parameters (e.g. MacKay, 1992; VanderPlas, 2014). The commonly used L1- and L2-norms directly correspond to a Bayesian Laplace or Gaussian prior, respectively.

Depending on the model type, models sometimes naturally contain constraints, e.g., by enforced physical principles like conservation laws. This additional information prevents such models to fit "non-sense" patterns in the data. For example, in hydrosystem models, mass balance prevents that fitted discharges or concentrations can obtain negative values due to their physical constraints. This can be considered as sort of model-type specific regularization in the sense of additional information.

Between models, complexity control is achieved by model rating and subsequent selection or combination (via averaging as discussed in Section 2.3). In order to account for structural deficiencies that lead to overfitting and underfitting of single models, competing models with the same target QoI but with different complexity even ought to be set up and tested. Between-model complexity control means, then, to rate these competitors under inclusion of a certain model complexity representation (law of parsimony), and elicit the model with the most appropriate complexity for the modelling task at hand. Based on the rating scores that the models achieve, a single model or model combination is found that resembles the appropriate complexity for the model task at hand.

Modellers typically refer to a rather vague notion of model complexity in the context of underfitting and overfitting. Höge et al. (2018) systematically analysed and discussed the model selection criteria from Section 2.5 with respect to their specific takes on model complexity. There, the explicit representation of model complexity within each class B1, B0, A1 or A0 conveys a distinct meaning. The decisive role thereof will be highlighted in the following. As in Höge et al. (2018), I will discuss it in the extremes of the $\mathcal{M}$-closed and $\mathcal{M}$-open setting, in the following also referred to as finite and infinite dimensional truth, respectively.

With the exception of Sections 3.2.2, 3.4.4 and the corresponding class-specific model complexity evaluations in Appendix B, the rest of the remaining Chapter 3 has been published in Höge et al. (2018) and I reuse parts of the text, figures and tables. Considering my co-authors, "I" is substituted by "we".

3.2 The Role of Model Complexity within Model Selection Criteria

All model selection methods that consider model complexity strike a trade-off between goodness-of-fit $-\mathcal{F}$ and model complexity $\mathcal{C}$ (e.g. Wit et al., 2012). In most general terms, there is a trade-off score $\mathcal{S}$ that is expressed as:

$$\mathcal{S} = -\mathcal{F} + \mathcal{C} \tag{40}$$

Traditionally, a model is rated better under a certain selection method, the more negative $\mathcal{S}$ is. This implies a good fit of data (hence the negative sign) and a low complexity (hence the positive sign). The goodness-of-fit $-\mathcal{F}$ is rather clear to interpret as the accuracy of the model, either based on a representative estimator like the maximum likelihood estimator (MLE), or based on an average fit, for example marginalized over the whole distribution of possible parameter values (van der Linde, 2012). Yet, the way model complexity $\mathcal{C}$ is interpreted and quantified differs strongly between model selection methods, as will be discussed in detail in the following sections.

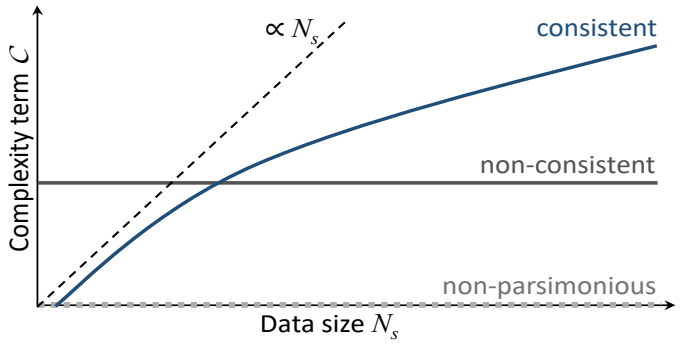


Figure 7: Schematic behaviour of complexity term $\mathcal{C}$ under growing data size N_s for non-parsimonious (light grey dots), non-consistent (grey line) and consistent (blue line) model selection with linear complexity growth as reference (dashed black line) (from Höge et al., 2018).

In non-consistent model selection, the complexity term is constant or bounded (Leeb and Pötscher, 2009), i.e. does not grow with the data size N_s. This is schematically depicted in Figure 7. Hence, non-consistent model selection allows switching to models of higher model complexity with more data as long as the higher complexity is compensated by an even stronger increase in goodness-of-fit. A special case of this would be non-parsimonious model selection where only the goodness-of-fit term $\mathcal{F}$ is used for rating models, and $\mathcal{F}$ is given by e.g. the maximum likelihood, smallest root-mean-square error or another error metric. This implies $\mathcal{C} = 0$ (see Figure 7) for all models and prevents generalizability or consis-

tency for the selected model because no trade-off is considered.

Opposed to this, the complexity representation in consistent model selection grows with increasing sample size N_s. However, this growth must be slower than linear (called subextensive and shown in Figure 7) (Bialek et al., 2001): mathematically, for growing data size $N_s \to \infty$ the complexity term follows $\mathcal{C} \to \infty$ and $\mathcal{C}/N_s \to 0$ (Leeb and Pötscher, 2009). Such growth might be contradictory to our intuitive understanding of complexity: If a model has a certain complexity, this complexity should not increase with increasing N_s. However, in consistent model selection, the model complexity penalty needs to grow in a way that the selection criterion can identify the true model, rather than justifying higher and higher model complexity with more and more data. While the goodness-of-fit will eventually get worse for all non-true model candidates with more data, only the true model can balance the growing complexity penalty.

3.2.1 Consistency in Model Selection

Accordingly, consistent model selection is sometimes also called confirmatory (Aho et al., 2014), i.e. confirming the identified DGP by the given data $\boldsymbol{D}$ in hindsight. Non-consistent model selection is also called conservative (Leeb and Pötscher, 2009) or exploratory (Aho et al., 2014), i.e. the model selected to approach the DGP is appropriate to conservatively predict or explore new data $\boldsymbol{D'}$ in foresight.

In the past, it was discussed whether the two types of model selection are (anti-) correlated (e.g. MacKay, 1992) or uncorrelated (e.g. Bishop, 1995) with each other. Although such behaviour might appear coincidently, it was generally shown that any model selection method cannot be optimal in both respects (Hurvich and Tsai, 1989; Yang, 2005; Arlot and Celisse, 2010).

Illustrative thought experiment
The exploratory or confirmatory natures of the two model selection types can be illustrated by a simple thought experiment: Imagine two modelers A and B who seek to model a controlled laboratory experiment (e.g. a tracer flow-through column experiment). Due to the fully controlled conditions it can be assumed that this lab-scale truth is of (relevant) finite dimensionality. Modeler A, e.g. an engineer or manager, assumes that there are too many dimensions to be covered by a fixed parametric model, but still wants to find the best model for future predictions. Accordingly, she picks a type of model which is allowed to grow with incoming new information and starts off with operational data-driven models, e.g. regressive models. Modeler B, e.g. a fundamental scientist, wants to identify the true data-generating process and hence prefers parametric physics-based models.

One might think that the two purposes are the same thing, but from the perspectives of non-consistent vs. consistent model selection, they are not.

Each of them starts with three models of their preferred model type with increasing complexity: A simple first model, a more complex second model and a highly complex third model. Let's assume that the second model of modeler B actually represents the truth (which is an idea borrowed from consistent selection), i.e. employs the right physical equations. On the same level of complexity, the second model of modeler A mimics the data best, but as a data-driven empirical model it is clear that it cannot represent the true data-generating process.

Both modelers collect and use the same data continuously in order to perform a model selection procedure as soon as a new batch of data, i.e. new and non-redundant information, comes in. According to her modelling purpose, modeler A uses a non-consistent model selection criterion targeting the highest predictive performance. Modeler B performs consistent model selection to identify the truth and to understand the underlying physics. This procedure is shown schematically in Figure 8.

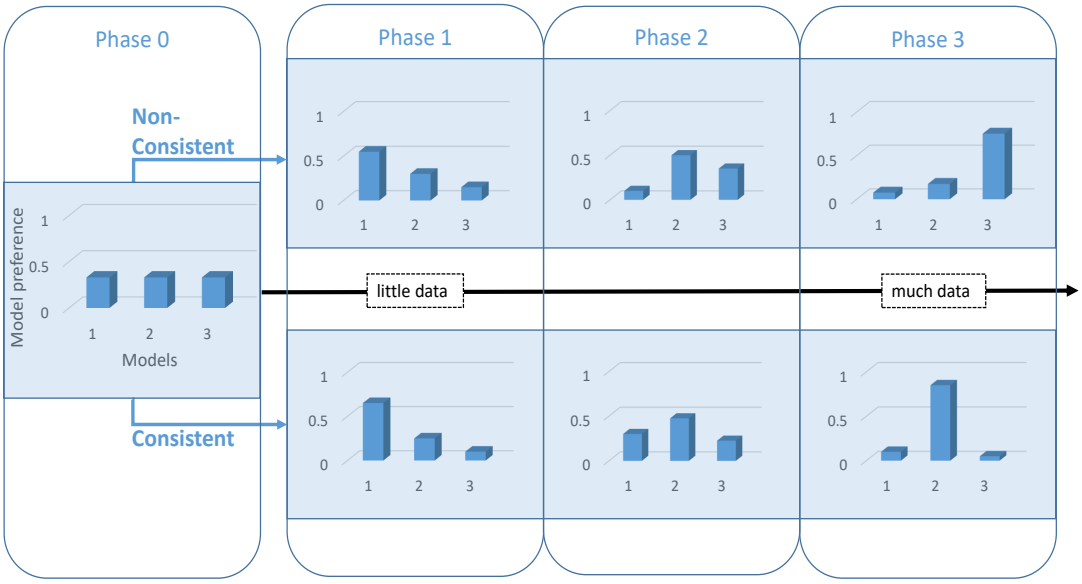


Figure 8: Differences in model rating following non-consistent (A-type) and consistent (B-type) model selection for increasing data size. The models are rated on a normalized scale between 0 and 1. Models 1, 2 and 3 resemble increasing stages of complexity (from Höge et al., 2018).

In Phase 0, before having any data, both modelers start with uniform model choice preferences across their candidate models. In Phase 1, with little data available, no complex model can be supported, so the simple first model of each modeler is selected. However, with more incoming and informative data (Phase 2), a more complex model provides a better trade-off between fit and complexity. Hence, the second models of both modelers get selected by their respective criteria. With more and more data becoming available in Phase 3, the two rankings become fundamentally different in the large sample limit: For modeler B the third physical model (which is more complex than the truth) will never stand a chance in a model selection process in the long run. Its additional complexity would be called excessive. However, the third data-driven model of modeler A can be justified as the model with the best trade-off between fit and complexity from an non-consistent perspective.

This is because, for modeler B, the second model revealed itself as representing the data-generating process, and as such a simpler (1st model) or more complex (3rd) model is rejected by the consistent model selection procedure. For modeler A it was clear from the beginning on that the truth is not among the data-driven candidate models. Then, a more complex model is justifiable with more available observations. More data reduces the risk of just fitting noise, so a more complex model from the efficiency perspective is confident with yielding the best future predictions and wins the model selection.

The illustrated behavior of consistent model selection, i.e. to identify and stick to the best representation of the truth, can be found in Schöniger et al. (2015a). In this study on mechanistic models for a laboratory-scale artificial aquifer, several increasingly complex parametrizations of the hydraulic conductivity distribution are ranked. Under growing data size, the consistent selection procedure converges towards the model that represents the true zonated distribution, and it devaluates simpler (homogeneous) and more complex (geostatistical) approaches. Contrarily, the tendency of non-consistent model selection to prefer increasingly complex models is demonstrated in Vrieze (2012) for regression models.

3.2.2 Bounds of Consistent Model Selection

Now, it is schematically clear how the consideration of model complexity $\mathcal{C}$ within model rating scores enforces consistent or non-consistent model selection. However, it is less apparent where the regimes are delimited from a statistical perspective: Under sub-extensive growth of $\mathcal{C}$ ($\mathcal{C} \to \infty$ and $\mathcal{C}/N_s \to 0$ for $N_s \to \infty$; e.g. Leeb and Pötscher, 2009) model selection is at least weakly consistent (Shibata, 1986), i.e., is per se able to identify a true model. Otherwise, for slower growth

of $\mathcal{C}$, it is non-consistent. Additionally, the Hannan-Quinn-Criterion (Hannan and Quinn, 1979) resembles the lower bound of strongly consistent selection behaviour (Shibata, 1986): any model selection method of which $\mathcal{C}$ grows faster than $\mathcal{C}_{\text{HQ}}$ converges to the true model almost surely in a statistical sense. Its complexity representation $\mathcal{C}_{\text{HQ}} = 2N_p\ln(\ln(N_s))$ is based on the so-called law of iterated logarithms (Shibata, 1986).

Interestingly, the law of iterated logarithms (LIL) is asymptotically (for $N_s \to \infty$) closely related to the two most famous limit theorems: the law of large numbers (LLN) and the central limit theorem (CLT). Such limit theorems are also known as universal laws that describe aggregate properties of a system that appear on a macroscopic scale despite a microscopic (or element-wise) complex and unpredictable nature (Tao, 2012). For a sum of N independent and identically distributed (i.i.d.) random variables: $S_x = x_1 + x_2 + ... + x_N$ (Klenke, 2013, and references therein),

- the LLN describes that the average of S_x converges to its expected value μ_x of X: $S_x/N \to \mu_x$ for $N \to \infty$;
- the CLT describes the typical fluctuations about the expected value that follow a normal distribution: $S_x/\sqrt{N} \to \mathcal{N}(\mu_x, \sigma_x^2)$ for $N \to \infty$; and
- the LIL describes the asymptotic behaviour of maximum deviation for S_x/N about the expected value μ_x, i.e. concerns the order of magnitude of (atypical) fluctuations: $S_x/\sqrt{2N\ln(\ln(N))} \to 1$ for $N \to \infty$.

In practical applications, the LIL is used, e.g., to mathematically describe the occurrence of sports records over time, when the typical results in the particular sports discipline follow a normal distribution (see Gembris et al., 2007).

In model evaluation and rating, the i.i.d. random numbers are the errors or residuals between model predictions and observed data. The universal laws underpin consistent model selection in the $\mathcal{M}$-closed setting, because only there the true model will yield i.i.d. residuals that follow these laws. For further details on the limit theorems, the interested reader may refer to, e.g., Klenke (2013); Tao (2012); Klesov (2014). Ties between the asymptotic behaviour of model selection criteria - or rather their model complexity representation - and such universal laws are evident and, interestingly, they mathematically delimit consistent model selection.

3.2.3 Bayesianism in Model Selection

While consistency refers to the goal of the model selection task at hand, Bayesianism refers to the statistical perspective used to achieve it. Many of the non-

consistent and consistent model selection methods are Bayesian to some degree. The Bayesian view allows assigning distributions to both data and parameters (Betancourt, 2015). Therefore, generally expressed, model selection methods which are Bayesian consider model complexity as "a measure for stochastic dependence between observations and parameters" (van der Linde, 2012). Model selection can be Bayesian in one or more of the following respects:

1. within-model expert knowledge:
   incorporation of prior probability distribution for parameters

2. model-representative quantities:
   measures of fit and complexity are marginalized over the probability distribution of the whole parameter space, and are not only (e.g. best-fit) point-estimates

3. between-models expert knowledge:
   model weights as prior ideas about ranking are used and updated, resembling model probabilities

The first level of Bayesianism in model selection addresses what the parameter space of a model looks like. In the Bayesian perspective, there is a probability measure (here represented for simplicity by a probability density function, pdf, $p(\boldsymbol{\Theta})$) of parameter values $\boldsymbol{\Theta}$ which expresses the belief what suitable parameter values could be. Even before observations are considered, there is such a belief and it is called the parameter prior pdf $p(\boldsymbol{\Theta})$. Observations $\boldsymbol{D}$ are then used to update this prior knowledge to a conditional distribution called the parameter posterior pdf $p(\boldsymbol{\Theta}|\boldsymbol{D})$.

The second level addresses whether a point (single parameter set) or a marginalized (averaged over the parameter pdf) estimator (van der Linde, 2012) shall be used to evaluate goodness-of-fit and model complexity. Often, goodness-of-fit is evaluated with the best parameter calibration possible for a model, i.e. with the maximum likelihood estimator (MLE). Being one specific set of parameters, the MLE is one of the classic point estimators. Contrarily, the fully Bayesian spirit is to marginalize over the whole parameter pdf (Piironen and Vehtari, 2017), and to use averaged quantities such as the marginal likelihood to represent the overall fit.

The third level refers two how models are compared, ranked or selected using model probabilities. From a Bayesian point of view, model selection is based on a belief in each model, again expressed as a probability P. There is a prior probability of each candidate model to be the model which has most likely generated the data $\boldsymbol{D}$. These data can then be used to update the prior model weights to their

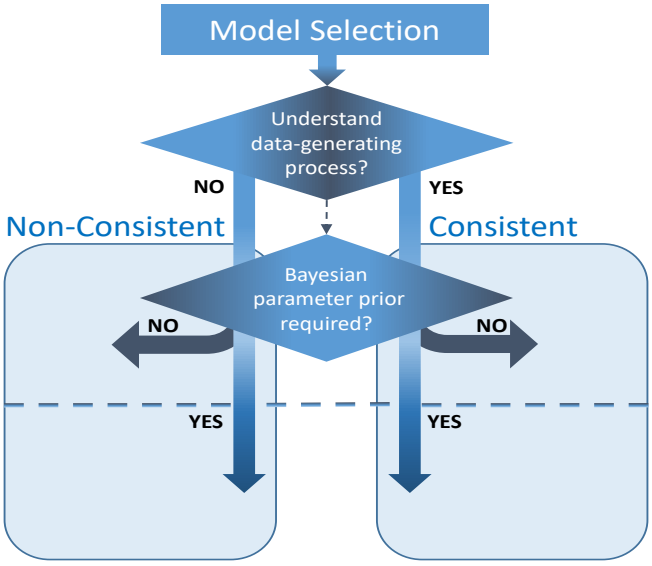


Figure 9: Classification system for model selection methods with 4 classes: 1st) Non-consistent versus consistent model selection, 2nd) Using a Bayesian parameter prior or not (from Höge et al., 2018).

respective posteriors, just as for parameters within the models.

For a model selection method to be Bayesian, at least the first level of Bayesianism has to be fulfilled. Figure 9 graphically summarizes the general classification scheme of model selection methods over the two stages covered in Sections 3.2.1 and 3.2.3. First, the non-consistent or consistent type has to be picked depending on the major purpose of modeling. Second, the incorporation of a Bayesian parameter prior (first level of Bayesianism) allows for a probabilistic treatment of parameters during the model selection task. The second and third level of Bayesianism are added on top of that by specific methods, as discussed for the respective methods in section 2.5.

3.2.4 The Role of Priors in Model Selection

Generally, the use of priors (for parameters and models) in model selection is a double-edged sword: On the one hand, an inappropriately chosen prior can yield problematic results or even allow a modeler to manipulate a model ranking in favour of a certain candidate model (Gelman et al., 2014). An appropriate prior that is too vague does not help either, preventing a clear model selection (Bartlett, 1957; Gelfand and Dey, 1994). The search for priors that are less susceptible to subjectivity of the modeler is still a large field of ongoing research. Among others, uniform, maximum entropy or reference priors (van der Linde, 2012) are investi-

gated as such "objective" priors.

On the other hand, using an appropriate (e.g. physics-based) parameter prior in consistent model selection might sometimes be the only way to get close to the data-generating truth. From a classic statistics point of view, infinitely many data points have to be collected until the best-fit parameter estimate of the true candidate model converges to the true parameters. In reality this is simply impossible, especially when expensive field data is collected. The parameter prior might be the missing piece to select the true model with limited data. Further, it serves as a natural regularization of the model that counteracts overfitting (VanderPlas, 2014). Hence, if for instance mechanistic models are used and a reasonable physical prior for the parameters is available, it shall be used (Vanpaemel, 2009).

3.3 Classification of Model Selection Criteria

The selection criteria presented in Section 2.5 are those most widely used by the majority of practitioners (see Mallick and Yi, 2013; Boisbunon et al., 2014). We classify and discuss them with respect to their incorporation of model complexity. Based on this classification, we explain why similarly looking but different selection methods yield contradicting model ratings (Ye et al., 2008; Schöniger et al., 2014) and put them into perspective, going beyond similar earlier attempts (e.g. McQuarrie and Tsai, 1998).

3.3.1 Classification Scheme

Choosing the right model selection class, as summarized in Figure 10, starts with asking what the purpose of the model is. This leads to either A-type (approaching truth) or B-type (identifying truth) model selection. Then, the next step refers to the distinction whether a Bayesian perspective starting with parameter prior incorporation shall/can be used or not.

Predictive information criteria (A1-type) are non-consistent, and also Bayesian to a certain degree: All of them cover the 1st level of Bayesianism; DIC and WAIC can incorporate informative priors. Even the AIC assumes a Bayesian parameter distribution, but just a non-informative one. Further, the WAIC uses averaged goodness-of-fit and model complexity terms (2nd level). However, none of these criteria is designed to work with Bayesian prior and posterior model weights (3rd level).

B1-type model selection is consistent and covers all three levels of Bayesianism. Methods of this kind use prior and posterior probabilities for both parameters and

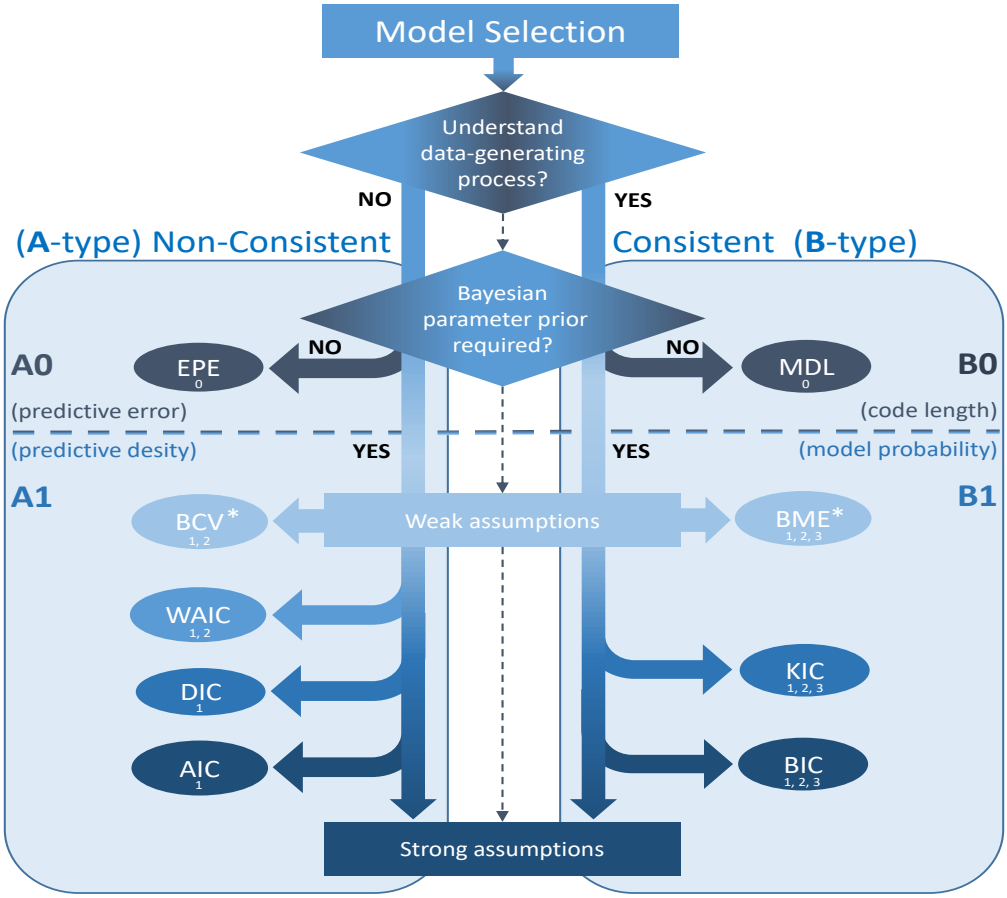


Figure 10: Classifying all introduced criteria based on them being 1st) non-consistent versus consistent and 2nd) Bayesian or not. Subscripts display the level of Bayesianism of the particular method/criterion. A1- & B1-type criteria are Bayesian using at least a Bayesian parameter prior. The corresponding criteria are sorted according to the strength of underlying assumptions in approaching their respective implicit (indicated by *) target model selection scores (BCV* vs. BME*). A0- & B0-type criteria are not Bayesian, they are represented by their most common members EPE and MDL, respectively. In the large sample limit $N_s \to \infty$, the influence of the Bayesian parameter prior declines and, respectively, non-consistent and consistent model selection criteria become asymptotically equivalent (dashed line) (from Höge et al., 2018).

models. Although point estimates are used in some of the criteria for assessing the goodness-of-fit, e.g. like in the KIC, they are used under conditions where they coincide with averaged estimators, i.e. the peak of a Gaussian likelihood function is both mean and maximum of the distribution. Further, although the 1st Bayesian level is covered by the BIC, it is irrelevant in the assumed infinite sample size limit.

The (weak-assumptions) end-members in A1-type and B1-type model selection methods are the implicit evaluation of a Bayesian cross-validation (BCV) score or

the Bayesian model evidence (BME) coming from a implicit evaluation, respectively.

Looking back at the three levels of Bayesianism, the third level (model probabilities/weights) occurs only in Bayesian model selection (BMS) (Hoeting et al., 1999). The other two levels may occur in both the non-consistent and the consistent model selection world. As an information-theoretic equivalent to Bayesian model weights, so-called Akaike weights (Burnham and Anderson, 2002, 2004) can be used in a similar way. However, these shall not be confused with the concept of Bayesian model weights, because Akaike weights are non-consistent and have no connection to the notion of (true) model probability.

While the Bayesian perspective is part of the underlying assumptions for A1-type and B1-type criteria, A0-type (e.g. EPE) as well as B0-type (e.g. MDL) selection criteria do not require a Bayesian parameter prior, as depicted in figure 10. They allow for prior-free non-consistent or consistent model selection, respectively. Hence, they are immune to misspecified priors, but can also not benefit from potential advantages of using a prior. However, this does not mean that they cannot be extended in a Bayesian fashion. For example, EPE can be employed with a Bayesian parameter prior as a form of regularization (Mallick and Yi, 2013). Similarly, the MDL (B0-type) presented here can be derived in a non-Bayesian context (Lanterman, 2001), but can be extended to the normalized maximum likelihood (NML) approach (Shiffrin et al., 2016) which is able to incorporate a Bayesian prior.

3.3.2 Contrasting the Views on Models and their Complexity

The four introduced model selection classes differ in the definition of models, the meaning of what a complex model is, how model complexity can be quantified and what the respective complexity measures are. Therefore, table 3 summarizes the foundations on which the four selection classes try to identify the respective "best" model.

A1-type criteria consider a model to be a probabilistic attempt to approach the infinitely complex data-generating truth - but only approaching, not representing. The best model achieves the highest predictive capability based on predictive density. A complex model shows a large offset (large N_p^*) between the estimated out-of-sample ($\boldsymbol{D}'$) predictive density and the within-sample ($\boldsymbol{D}$) predictive density (Vehtari and Ojanen, 2012), because the complexity of the model does not allow the data $\boldsymbol{D}$ to sufficiently constrain the parameters.

Table 3: Class-specific consideration of what models are in principle, what a best model provides and based on which score this is measured; respective properties of complex models, how complexity is represented and quantified, and which model selection criteria work accordingly (from Höge et al., 2018).

Type	Model is a ...	Best model...	... based on...	A complex model...	Complexity...	...quantifies...	Criteria
A1	probabilistic attempt to approach truth.	has largest predictive capability	predictive density.	has low predictive coverage.	N_p^*	data-constrained parameter number.	AIC, DIC, WAIC (, BCV)
A0	flexible regression of data.	poses most stable inversion	predictive error.	poses a non-unique inversion problem.	DoF	sensitivity to data perturbations.	EPE, Mallows' C_p
B1	probabilistic attempt to represent truth.	is most likely data-generating process	model probability.	allows only weak parameter inference.	OF	"posterior-prior -ratio".	BIC, KIC (, BME)
B0	compression of data series.	is most compact data representation	code length.	is a too long code.	GC	distinguishable likelihoods.	MDL

In a similar spirit, A0-type criteria assume that a model is just a more or less flexible regression of data. This does not need to be inspired by the physical truth behind the data, either. The best model obtains the highest predictive capability based on its predictive error for $\boldsymbol{D}'$. A0-type criteria are concerned with the instability (flexibility) of the model inversion. A complex model only allows instable or non-unique parameter inversion, which is measured by large sensitivities (large DoF) of model predictions with respect to perturbations in the data $\boldsymbol{D}$.

B1-type criteria take each model as a probabilistic attempt to truly represent the data-generating process, believing that the true model exists and is among the candidate models. The best model is most likely to have generated the data $\boldsymbol{D}$ and achieves the highest probability of being the true model. B1-type criteria expect the strongest parameter inference for this model and its prior when faced with the data $\boldsymbol{D}$. A complex model shows weak parameter identifiablity (large Occam factor OF), quantified as the shrinkage ratio from the prior towards the posterior parameter distribution.

Alternatively, B0-type criteria consider each model to be a compression of data. Thus, they state that the best possible compression of data requires just a certain code length. The best model is the most compact one, which according to coding theory coincides with the data-generating truth. Compactness of a model is quantified as number of distinguishable (likelihood) distributions over its parameter space. A complex model in a B0-type sense is a too long compression of $\boldsymbol{D}$.

3.3.3 Matching Model Selection Classes with Model Types

The choice of a certain model selection criterion is specific to the model selection task at hand. We outline imaginable extreme cases of matchings for the field of water resources in the following, but there are equivalents in practically all other fields where mathematical or numerical models are employed.

For A-type model selection in an infinite (relevant) dimensional truth scenario, matching suggestions between selection criteria purpose and models one could think of are:

- A1-type (N_p^*): Providing high predictive capability via high predictive density for unseen data. The probabilistic nature allows for incorporating prior parameter knowledge - example: Bucket-type models for stream discharge or flood forecasting (e.g. Orth et al., 2015). Such models normally include (semi-)physical relationships and corresponding prior parameter distributions.

- A0-type (DoF): Obtaining the highest predictive capability in a non-probabilistic manner via predictive error for unseen data with the most uniquely calibrated model. Besides, the effect of regularizations can be assessed in a second step because reduction in DoF means reduced risk of overfitting - example: Regression models (e.g. artificial neural networks) for time-series predictions (e.g. Tibshirani, 2014). Such flexible models are hardly uniquely calibrated and often require some sort of regularization.

For B-type model selection in a finite (relevant) dimensional truth scenario, matching suggestions between selection criteria purpose and models one could think of are:

- B1-type (OF): Identifying the data-generating model via Bayesian model selection, given that a reasonable prior is provided (e.g. based on physical quantities) - example: Partial differential equations (pde)-based models for groundwater flow (e.g. von Gunten et al., 2014). These mechanistic models and their parameters are subject to prior knowledge and physical meaning. However, it is crucial that prior information on parameters covers potential subsurface heterogeneity, scale-dependence, etc. to obtain a maximally unbiased prior (e.g. for hydraulic conductivity).

- B0-type (GC): Approaching the true model via minimal required code length, without the need to specify a certain parameter prior distribution - example: Stochastic rainfall generators (e.g. Golder et al., 2014). Such a model represents the statistics of the process of interest. The parameters are the statistical moments which describe the process pattern.

These extreme examples of matching highlight the importance of having the model purpose, the type of model (data-driven,mechanistic,...), and the information about the model parameters in mind (Guthke, 2017), when an appropriate model selection class has to be picked. Further, the consideration of the (relevant) dimensionality of the truth to be modelled affects whether a certain kind of model selection matches with a model (Leeb and Pötscher, 2009).

3.3.4 Alternative Model Selection Criteria

Apart from the model selection criteria presented above (AIC, BIC,...), many other criteria were developed over the last decades - to the point that nearly a whole alphabet of criteria can be set up (Spiegelhalter et al., 2014). In most of them, model complexity is interpreted and measured differently, some are advances or refinements of other criteria. Additional examples of widely used model

selection criteria and complexity measures are the non-consistent ICOMP (Bozdogan, 1990, 2000), Moody's effective number of parameters (Moody et al., 1992) or the Vapnik-Chervonenkis (VC) dimension (e.g. Friedman et al., 2001) which is used in structural risk minimization (Guyon et al., 2010). Additional consistent model selection criteria are Hannan-Quinn (Hannan and Quinn, 1979) or various versions of encoding complexity (Rissanen, 1987; Myung et al., 2006).

Covering all of these in detail would go beyond the purpose of this primer, but the completed classification scheme in section 3.3.1 may allocate them as well. Crucial in their application is always how exactly they consider model complexity.

3.4 Cross-Comparison between Model Selection Criteria

3.4.1 B1- vs. B0-type Criteria

Interestingly, the integrand from GC in equation 29 coincides with the so-called Jeffrey's prior for parameters $p(\boldsymbol{\Theta}) = \sqrt{|\mathbf{F}(\boldsymbol{\Theta})|}$ (Myung et al., 2006). This prior is used as a kind of non-informative or "objective" prior in Bayesian model selection (Barron et al., 1998). For large N_s and using Jeffrey's prior on parameters, BMS is identical to model selection using MDL (Myung et al., 2006). Further, in the limit of $N_s \to \infty$, the last term of MDL in equation 29 becomes negligible due to its independence on sample size. Further, in this limit, the prior model weights and prior parameter distribution become irrelevant. Then, MDL becomes proportional to the BIC because the complexity terms in both criteria scale equivalently with $\ln N_s$ and N_p (Barron and Cover, 1991; Myung et al., 2000; Hansen and Yu, 2001; Shiffrin et al., 2016).

Despite the asymptotic equivalence, the criteria and complexity representations differ fundamentally between the two classes (B1 & B0) as depicted in Figure 10: Model complexity in B1-type criteria (OF) relates to how much knowledge about the parameters was inferred from data $\boldsymbol{D}$, shrinking the prior to the posterior distribution. A complex model in this sense is a model for which parameters can hardly be constrained and identified with $\boldsymbol{D}$. B0-type model complexity according to coding theory is measured as GC and relates to compressibility of data. A complex model in this sense is a long code needed to describe the regularities of data $\boldsymbol{D}$. Apart from the special cases above, real BMS requires using posterior model weights, Bayesian parameter and model distributions, all of which is not supported by MDL. Therefore, the two classes generally lead to different selection results (Grünwald, 2000), unless the true model is actually among the candidates and will eventually be selected. Overall, the common ground of consistent model selection (shown in B1 & B0) can be summarized by three points:

1. Model complexity is a measure for the lack of identifiability of a model and its parameters as representation of the data-generating process.

2. Model complexity is an integrated quantity over all possible model parametrizations unconditional on data $\boldsymbol{D}$. Consistent model selection compares model predictions to data $\boldsymbol{D}$ over the whole parameter space.

3. Their behaviour in the limit of $N_s \to \infty$ is asymptotically equivalent.

3.4.2 A1- vs. A0-type Criteria

For linear models and uncorrelated Gaussian errors, it can be shown that the A0-type EPE (equation 38) is equivalent to the A1-type AIC (equation 31) (Mallick and Yi, 2013; Boisbunon et al., 2014). In this special case, the model complexity representation by N_p coincides in both criteria. This coincidence is one of the reasons, why model DoF and N_p^* are often used interchangeably to quantify model complexity despite their different motivation. In classical statistics, DoF are a measure for the "number of dimensions in which a random vector may vary" (Janson et al., 2015). This interpretation also suits the flexible-parts-view on DoF and triggers even more the interchangeable use with effective number of parameters N_p^*. Further, in the large sample limit $N_s \to \infty$, the influence of the Bayesian parameter prior in the DIC and WAIC declines. This makes A1-type criteria asymptotically equivalent to A0-type criteria.

Despite these similarities, they are not the same thing: A0-type DoF in model selection refer to the difference between two kinds of error (training & test). They resemble the summed-up sensitivities of predictions to perturbations in the calibration data and can be seen as quantifying the stability of the model inversion. A1-type N_p^* refer to the information-theoretically motivated distance between probability distributions of observations and model predictions, which requires the incorporation of a Bayesian parameter prior. Due to their different motivations, they correspond to different non-consistent model selection classes, as shown in Figure 10. Overall, the common ground of non-consistent model selection (shown in A1 & A0) can be summarized by three points:

1. Model complexity is an estimate for the lack of generalizability to unseen data $\boldsymbol{D'}$ after seeing $\boldsymbol{D}$. In a quite counter-intuitive manner, this is estimated based on just the calibration data $\boldsymbol{D}$ (in IC), i.e. without actually considering a validation data set $\boldsymbol{D'}$ (as in CV).

2. Model complexity is evaluated for the model having a certain parametrization (calibration) conditional on data $\boldsymbol{D}$.

3. Their behaviour in the limit of $N_s \to \infty$ is asymptotically equivalent.

3.4.3 A-type vs. B-type: Large Sample Limit

In the limit of infinitely large sample size $N_s \to \infty$, the parameter prior distribution becomes negligible and the complexity terms of the criteria within the A-type classes converge as well as those within the B-type classes. However, non-consistent model selection differs fundamentally from consistent model selection especially in this limit (as schematically depicted in Figure 8 of the illustrative thought experiment). The criteria designed for this limit are the AIC and BIC, respectively. This is why model selection criteria are often sorted into the so-called AIC-world and BIC-world (Vrieze, 2012; Aho et al., 2014), which is conform with A-type and B-type used in this primer. The respective model complexity terms are shown in Figure 11 in order to visualize the fundamental difference between the two worlds. Remember, that the two criteria were designed for the large sample limit. Nontheless, AIC and BIC are displayed for small sample sizes in Figure 11 for two reasons: First, they are often applied in practice regardless of this assumption. Second, these prominent members of the two model selection types are perfectly suited to display the deviating model complexity representations between non-consistent and consistent criteria and what this implies for the selection of models.

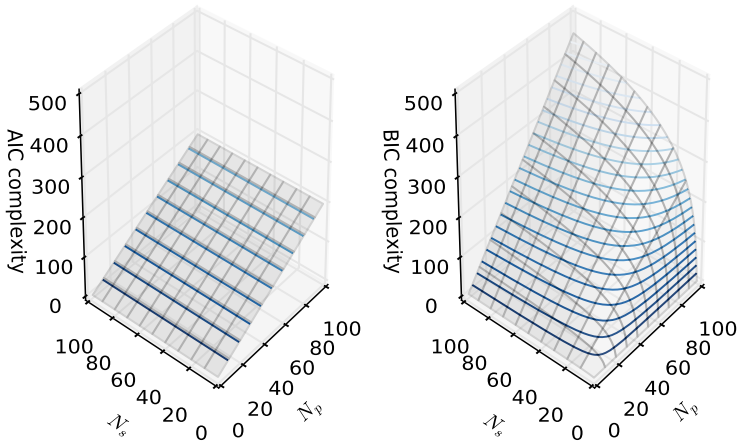


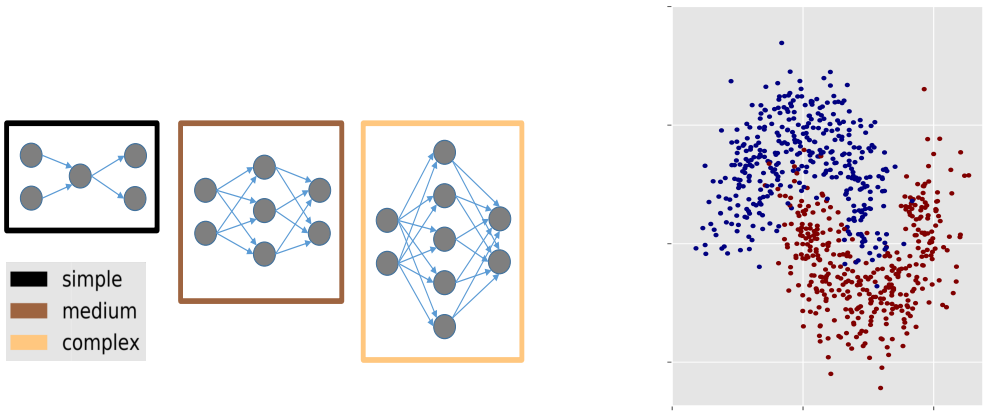
Figure 11: Model complexity representation of AIC (twice the number of parameters: $2N_p$) versus BIC (the number of parameters scaled by the logarithmic number of observations $\ln(N_s)\,N_p$). Blue isolines display same complexity in the $N_s - N_p-$ space, with lighter blue indicating higher complexity (penalty) (from Höge et al., 2018).

In the AIC-world, the complexity penalty generally does not grow with growing

N_s as can be seen in Figure 11. In the AIC, complexity is totally independent of N_s and is given as twice the number of parameters. This resembles the most classic way of bounded complexity representation (Leeb and Pötscher, 2009) in non-consistent model selection, enabling A-type criteria to successively approach the (infinite-dimensional) truth by switching to "closer" models under growing data size. Opposed to this, in the BIC-world, the complexity penalty constantly increases with $N_p \ln(N_s)$ as shown in Figure 11. This depicts the consistent nature of model selection criteria in the BIC-world in the simplest way, enabling these criteria to identify the (finite-dimensional) true model that is assumed to be among the model candidates.

3.4.4 Model Selection by AIC and BIC Exemplified

In a simple example, I want to demonstrate how the most popular ICs from the two parsimonious model selection worlds, AIC and BIC, are typically employed in practice. Therefore, both criteria are evaluated over growing data size. The feature application is the binary classification problem from Figure 6 by data-driven neural networks (NN; see Equation 3; built according to the tutorial in Britz (2015)) of different complexities (Figure 12 (a)). These are used to mimic the separating curve for classifying subsets of data that are drawn randomly from a total of 750 data points as to see in Figure 12 (b).



(a) **(b)**

Figure 12: (a) Neural networks (NN) of increasing complexity: simple (black box, left), medium (brown box, center) and complex (beige box, right); (b) Total data comprising 750 points with predefined blue and red labels in binary classification with unspecified attributes on the axes.

The loss function for this classification with $N_L = 2$ labels is the so-called cross-

entropy between the true labels $\boldsymbol{D}$ (red or blue) and the predicted labels $\hat{\boldsymbol{y}}$ over N_s data points:

$$\mathrm{L}(\boldsymbol{D}, \hat{\boldsymbol{y}}) = -\frac{1}{N_s} \sum_{o=1}^{N_s} \sum_{l=1}^{N_L} D_{o,l} \ln \hat{y}_{o,l} \tag{41}$$

The cross-entropy in this discrete form is a typical likelihood function in machine learning, where it is used to rate the quality of classification models (e.g. Friedman et al., 2001). Generally, between the distribution of a true model $q(\boldsymbol{y}|M_{\text{true}})$ and a predictive model $p(\boldsymbol{y})$ the cross-entropy is typically written as $H(q(\boldsymbol{y}|M_{\text{true}}), p(\boldsymbol{y})) = E_{q_{\text{true}}}[\ln p(\boldsymbol{y})]$. It is the sum of the entropy of the true model $q(\boldsymbol{y}|M_{\text{true}})$ and the D_{KL} of $q(\boldsymbol{y}|M_{\text{true}})$ coming from $p(\boldsymbol{y})$ (cf. Section 2.2.4).

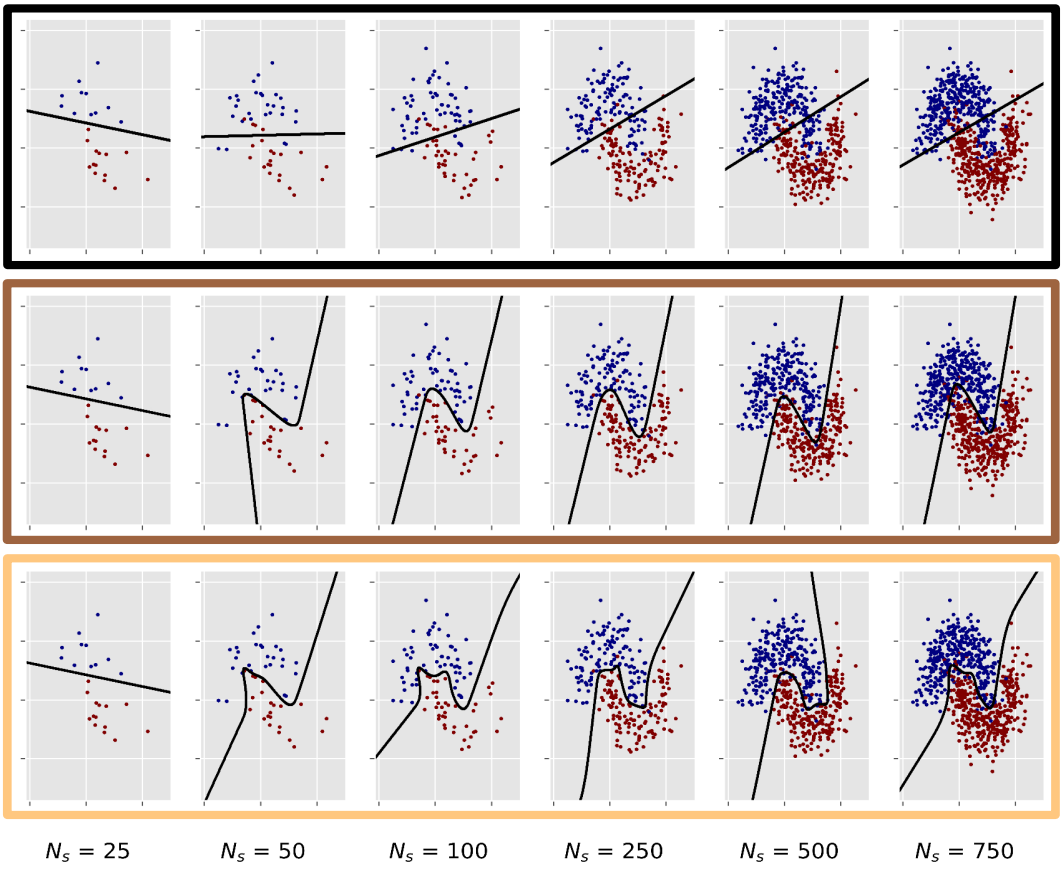


Figure 13: Maximum likelihood separation line in binary classification over growing data size N_s of the simple NN (black box, top), medium complex (brown box, center) and complex NN (beige box, bottom).

The model preferences of the non-consistent AIC and of the consistent BIC are evaluated over subsets of the data with $N_s = 25, 50, 100, 250, 500$ or 750. Figure 13 shows the maximum likelihood estimators (MLE) of separation lines for each of the three NN classification models at every subset.

Figure 14 shows the resulting model weights for the three models assigned by (a) the AIC and (b) the BIC. The AIC clearly shows its tendency to prefer more complex models when there is enough data to support it. Contrarily, the BIC depicts conservative model choice in search for the true model - which in this case is, however, not among the candidate models. From the AIC perspective of predictive model choice, it is legitimate to switch towards the medium-complexity model earlier than from the BIC perspective. As soon as enough data supports it, the AIC even switches to the most complex model. Sooner or later, for both criteria, the most simple model is considered to underfit and therefore not preferred. While the BIC then sticks to the medium-complexity model which it identifies as true one, the AIC moves on to the most complex model. While the flexibility of the separation-line from the most complex model is interpreted as too high complexity with respect to the true model by the BIC, the AIC supports this flexibility as legitimate approximation of an unknown truth given the large amount of available information ($N_s = 750$).

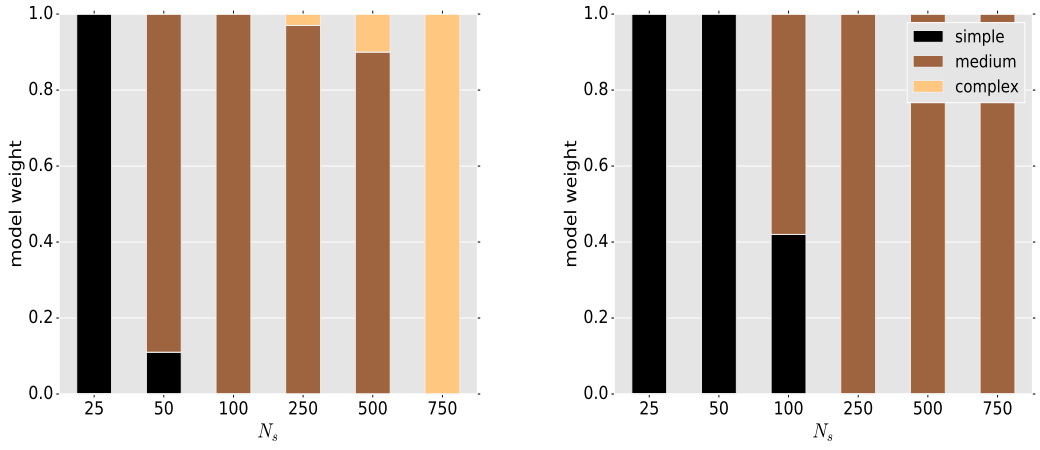


Figure 14: Model weights estimated by AIC (a) and BIC (b) in binary classification over growing data size N_s represented as fractions from bars of length 1 of the simple NN (black parts), medium complex (brown parts) and complex NN (beige parts).

This simple illustration highlights three important points:

- Whenever a multi-model framework is applied, it is crucial to keep in mind how the (explicit or implicit) complexity representation in the chosen score evolves over growing data size (as depicted in Figure 11). This directly pertains to how a method rates each model.

- Both, the AIC and BIC, were derived in the large sample limit. Therefore, their application for small data set sizes is questionable. The illustrative example refers to a comparably simple modelling task and therefore both criteria yield plausible results. However, in more demanding modelling tasks, violated assumptions might trigger wrong conclusions.

- While formally conducted correctly by using the same MLE, the AIC and the BIC still represent fundamentally different takes on model rating. Friedman et al. (2001) argue that the assigned likelihood function for misclassification does not fit to the context of BIC, i.e., consistent model selection. The three NN models that are rated did not generate the data-points but predict curves to separate labels. Under these conditions, the intention of the BIC to identify the model that most likely generated the available data in an $\mathcal{M}$-closed setting is inadequate. Hence non-consistent methods like the AIC should be employ. This highlights, again, that the adequacy of a certain model rating method is indispensably tied to the $\mathcal{M}$-setting of the modelling task.

3.5 Summary and Conclusion

Model selection methods perform an explicit or implicit trade-off between goodness-of-fit with data and model complexity. Generally, no complexity metric in model selection works without incorporating data - which means that there is no unique intrinsic model complexity (e.g. Du, 2016) that quantifies complexity only based on the model's functional relationships and parameters. The counted number of parameters N_p fully represents the complexity of the model only in special cases of model selection (see A-type).

It is non-intuitive why the two major model selection types (non-consistent and consistent) should not lead to selecting the same model. However, they are optimal under different assumptions about the dimensionality of the truth that is modelled. If this truth is infinite dimensional, a model selection method is optimal if it can progressively approach this truth by sticking with one model only until more data justifies switching to another (more complex) one that approaches the truth even more closely (A-type model selection). Alternatively, if the truth is of finite (relevant) dimensionality, a model selection method is optimal if it identifies the

model that fully parametrizes this truth (B-type model selection). Hence, both types of model selection pursue different target quantities for model selection and yield deviating results when they are applied to the same modeling task.

The model purpose is crucial to be considered when a particular model selection method is used. From a pragmatic point of view, non-consistent model selection is the right choice for finding the best model for predictions in situations where the modeller cannot be sure that the truth can be sufficiently represented. Then, non-consistent methods enable optimal use of a certain model until more observations become available and a more complex model can be legitimately employed. Driven by the philosophy to find the model which represents the truth, a model selected in a consistent manner will avoid to be falsified when more data arrives. The consistent selection therefore ranks candidate models (hypotheses) according to how strongly they resist to be proofed wrong by the data. Therefore, consistent model selection is the right choice for process understanding and scientific hypotheses testing because it is philosophically completely in line with the scientific approach.

Centered around the specific interpretations of model complexity, we conclude the following major points:

1. When choosing between model selection criteria, the truth (dimensionality) that shall be approached or represented by a certain type of model indidcates the appropriate type of model selection. Whether this modeling purpose can be pursued in an either Bayesian way or not, directs towards the right model selection class. The assumptions met by the modelling task at hand justify the corresponding method/criterion within each class.

2. Model selection methods that incorporate Bayesian priors should only be applied if "reasonable" priors can be assigned. The purpose of the prior should be to provide a meaningful context for testing models (Nearing and Gupta, 2015), which means not too vague and not too constraint in order to allow for a fair model selection. In cases where a "reasonable" prior cannot be assigned, non-Bayesian model selection methods offer an alternative.

3. Some of the *explicit* model selection criteria underlie strong assumptions in order to reliably quantify what they consider to be model complexity. If these assumptions do not hold, we rather recommend an admittedly more computationally costly but more reliable *implicit* method, e.g. (Bayesian) cross-validation (non-consistent) or direct evaluation of Bayesian model evidence (BME) (consistent).

4. For general discussions during qualitative model development and comparison, it does not seem to be necessary to force our intuitive notion of complexity into a specific definition - which certainly will not be comprehensive and itself will be subject to discussion (see Gell-Mann, 1995). However, as soon as a model selection technique is applied, a specific definition and role of model complexity is used and the models are ranked accordingly. A comparison of different model selection metrics does therefore only make sense if either they belong to the same class (e.g. B1) or if their respective interpretation of model complexity is part of the discussion on the results.

5. Rather than claiming the "best" model was found with a certain model selection criterion, it would be more appropriate to call it "best given the complexity interpretation" of the particular criterion. All of the criteria give the right answer (within their underlying assumptions and limitations), but to different questions.

4 Applying Bayesian Multi-Model Frameworks Properly to Model Settings

Modelling in practice normally does not offer the obvious and clearly distinguishable $\mathcal{M}$-settings (finite vs. infinite dimensional true model) as discussed in Chapter 3, with their clear prescription for model rating and selection. In applied modelling, we employ multi-model frameworks to cope with the conceptual model uncertainty under typically vague conditions. Sometimes, it is difficult enough to develop a single model that is plausible. Then, by expanding, reducing or varying certain model parts of this single model, it is usually possible to generate other model candidates in case conceptual alternatives were not present right from the beginning. Either way, finally having an ensemble of models available, its effective usage is complicated by mainly two issues:

- Having only limited observed data for inference and model rating.
- Allocating the task and models to the appropriate $\mathcal{M}$-setting.

For example, in hydrosystem modelling, the collection of field data is usually demanding and expensive. The system under investigation can hardly, if not impossibly, be fully investigated. Placing corresponding models in a certain $\mathcal{M}$-setting is therefore challenging and explanatory or predictive model rating might be contradicting goals.

As discussed for the three Bayesian multi-model frameworks in Section 2.3, model weights in different frameworks carry different meanings according to their underlying theory. Hence, I investigate the impact of both issues by applying the three frameworks to models for a typical modelling task in hydrosystem modelling that I introduce in Section 4.1. I relate this task to different $\mathcal{M}$-settings and propose a novel Quasi-$\mathcal{M}$-closed setting for applied modelling. Accounting for limited data in predictive model rating, I apply the Bayesian multi-model frameworks and the Bayesian Bootstrap (BB; see Section 2.4.2) within respective settings in Section 4.2. I compare the results of each method to contrast their outcomes in Section 4.3. Thereby, I specifically focus on the results obtained when frameworks are applied to settings they are not tailored for, e.g., BMS/BMA outside of $\mathcal{M}$-closed. I close the chapter with conclusions in Section 4.4.

4.1 Modelling Task, Data and Models

The modelling task, data and models chosen to demonstrate the three Bayesian multi-model frameworks from Section 2.3 were previously used and presented in

Schöniger et al. (2015a). There, the modelling task was to find the best out of several models that represent different subsurface parametrization of hydraulic conductivity for a laboratory-scale sandbox. The rating of models was done via BMS/BMA, based on drawdown data from a hydraulic tomography experiment in the sandbox studied by Illman et al. (2010).

4.1.1 Lab-scale Experiment, Data and Likelihood Function

The lab-scale sandbox is depicted in Figure 15 (cf. Illman et al., 2010). It is 193 cm long, 82.6 cm high and 10.2 cm deep, and it contains a synthetic heterogeneous aquifer made of eleven basic sand types. With these base types and mixtures, 18 layers were created by cyclic deposition of sediments (Illman et al., 2010). During this process water flow rates and feed rates of sediment were varied in order to mimic natural sedimentation. The photograph in Figure 15 shows the interfingering layers of the sandbox. 48 ports, each with a diameter of 1.3 cm, are available to access the sandbox aquifer for pumping or measurement devices. Geologic core measurements from these ports define the range of hydraulic conductivity K values to be log-normally distributed with a unit-normalized mean of $\overline{\ln K} = -2.56$ ($K = 0.077$cm/s) and a variance of $\sigma^2_{\ln K} = 0.87$ (Schöniger et al., 2015a).

Steady-state drawdown observations serve as data. From the 48 ports, only 36 were used to collect data, 12 were discarded due to a bad signal-to-noise ratio (Schöniger et al., 2015a). Throughout 6 pumping tests, in total 210 observations were collected as data $\boldsymbol{D}$ - using one port at a time for pumping and the others to measure drawdown yielding 35 observations per pumping test. These ports were selected regularly distributed over the whole sandbox and the resulting $\boldsymbol{D}$ served as calibration data for parameter posterior inference. Four additional pumping tests were conducted using ports that served only as observation ports before. This produced another 140 observations as validation data $\boldsymbol{D}'$. Measurement errors were assumed to be uncorrelated Gaussian with a standard deviation of 1 cm. This is represented in a Gaussian likelihood function centred at the data $\boldsymbol{D}$ for calibration and $\boldsymbol{D}'$ for validation, respectively.

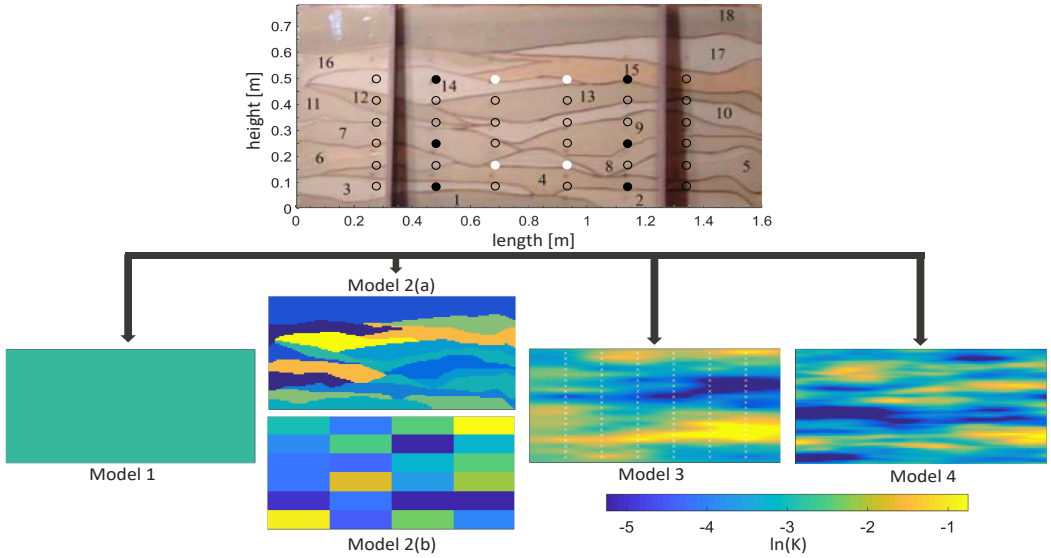


Figure 15: Top: Photograph of the laboratory sandbox aquifer (modified from Illman et al., 2010). The different coloured layers resemble different hydraulic conductivities, numbers enumerate the layers. Circles indicate the locations of the ports - for pumping to obtain calibration (black) and validation (white) data. Bottom: Corresponding numeric models (modified from Schöniger et al., 2015a): Homogeneous (1), informed zonated (2a) and uninformed zonated (2b), pilot-point-based (3) and geostatistical (4).

4.1.2 Mechanistic models

For simulation, the sandbox was discretized in a physics-based finite element (white-box) model based on partial differential equations for steady-state groundwater flow (see Equation 2 without storage term). The models were implemented as two-dimensional over a window of 160 cm length and 78 cm height in order to avoid boundary effects (Schöniger et al., 2015a). The PDEs were solved on a regular grid with a spatial resolution of 1 cm in each direction, resulting in 12,480 cells with 12,719 nodes. The forward model runs were then executed using a vectorized FEM solver (Nowak et al., 2008), which employs the standard Galerkin technique for spatial discretization (Cirpka and Nowak, 2004). For the 2-D sandbox model, the FEM elements become a regular grid of square elements. According to the chosen parametrization, each element is assigned a certain hydraulic conductivity value. Within each element, e.g. at the exact port locations, hydraulic heads are obtained by bilinear interpolation which is numerically consistent with the FEM interpolation. Each forward model run returns the parameter-specific model forecasts for steady-state hydraulic heads at the observation ports.

All parametrizations of the 2-D model were based on the physical properties of the sandbox (mean, variance, and correlation lengths). These subsurface parametri-

zations as shown in Figure 15 depict a representative range of model complexity within physics-based groundwater models which are often applied to larger real-world scales:

1. a homogeneous single effective parameter model (1 parameter)
2. a zonated model
   (a) with an informed zonation model inspired by the spatial distribution of sand layers (19 parameters)
   (b) with an uninformed zonation model that accounts for parameter heterogeneity but without realistic spatial distribution (24 parameters)
3. a deterministic geostatistical model by kriging based on stochastically parametrized pilot points for $\ln K$ (120 parameters)
4. a stochastic geostatistical model generated from Fast Fourier Transform-based logarithmic multi-Gaussian random fields (12480 parameters)

These models are very similar to so-called nested models, i.e., model alternatives of which simpler models are contained in the more complex ones, making them generalizations.

In order to assure full sampling of the model-specific prior parameter distributions (cf. Section 2.4.2), large parameter ensembles were generated by brute-force Monte Carlo sampling of each model's $p(\boldsymbol{\Theta}|M_m)$. These ensembles comprise of $2.0 \cdot 10^5$ samples for the homogeneous model and $1.0 \cdot 10^7$ samples for the other four approaches. The large ensemble sizes assure convergence of inferred posterior distributions and of the evaluated model rating scores.

4.1.3 Summary of the Reference Study

The reference study by Schöniger et al. (2015a) focused exclusively on the BMS/BMA framework for model rating:

- Over growing size of data $\boldsymbol{D}$ from successive pumping test inclusion, BMS/BMA was conducted and it was shown how the results at each stage can be interpreted and utilized.
- In a model justifiability analysis so-called model confusion matrices were introduced and demonstrated: In the large sample limit, BMS/BMA is supposed to converge to a weight of one for the allegedly true model. Using only a limited amount of observed data, a justifiability analysis yields the

maximally expectable model weights in model selection which are presented as confusion matrix.

- In a contrasting example, the uninformed zonated model (2b) was included into the model set as substitution for the informed zonated model (2a). Then, BMS/BMA was applied to and interpreted for this modified model set.
- In a validation, the predictive performance was tested on hold-out validation data which had not been used for the Bayesian inference.

BMS/BMA was evaluated for 6 data set sizes, i.e., with one to six included pumping test (PT) and with 35 measurements per pumping test. At each stage of included pumping tests, all possible combinations of pumping tests were evaluated and the resulting BMS/BMA-weights were averaged: When 5 from the total of 6 pumping tests are considered, 6 combinations are possible, 4 out of 6 yields 15 combinations, 3 out of 6 yields 20 combinations and 2 out of 6 yields 15 combinations. 1 or 6 out of 6 are trivial cases. Hence, in total, 63 combinations of pumping tests were evaluated and the behaviour of averaged model weights was analysed. The tendency of BMS/BMA to converge to the allegedly true model (informed zonated, 2a) could thereby clearly be demonstrated.

The focus of Schöniger et al. (2015a) was to thoroughly investigate how strongly the proposed models are able to recognize themselves as DGP given the data size at each step of included number of pumping tests. Stepwise, the results were presented as a so-called confusion matrix: A confusion matrix displays the degree of each model in the ensemble to recognize itself or any alternative model as the data-generating model given the current amount of data by respective model weights. It was built from generating 1000 model outputs as synthetic datasets $\boldsymbol{D}^{syn}$ from each model in the ensemble. Then, all models were conditioned on the synthetic datasets from one another and, for each dataset $\boldsymbol{D}^{syn}$, corresponding BME values were obtained. Based on this, BMS/BMA was conducted and the resulting 1000 model weights of each model were averaged. In the sandbox case study, the confusion matrix for model 1, 2a, 3 and 4 showed that using the calibration data from 6 pumping tests (210 observations) does not allow for full identification of the informed zonated model (2a) as true model even if it actually generated $\boldsymbol{D}^{syn}$. Using these 210 data points, an averaged model weight of only 75% was achieved. Generally, Schöniger et al. (2015a) recommend to perform such a justifiability analysis before including the actual observations $\boldsymbol{D}$ to estimate the maximal model weight that can be expected by BMS/BMA given the size of $\boldsymbol{D}$.

Referring to real-world scenarios where knowledge about the modelled system might not be as highly resolved as on the lab-scale, the rougher, uninformed zo-

nated model (2b) was added to the ensemble in exchange for the informed zonated model (2a). Model 2b performed clearly worse than model 2a and therefore BMS/BMA did not converge to it anymore over growing data size, but rather to the pilot-point model 3.

The predictive performance of each single model and the BMA pdf-average for the model set (incl. model 2a) were evaluated using the the validation data $\boldsymbol{D}'$ (140 observations) via:

- the root-mean-square error (RMSE) between $\boldsymbol{D}'$ and the expectation over posterior predictions $\boldsymbol{y}'$ (see Equation 12) as a measure of accuracy:

$$\text{RMSE} = \sqrt{\frac{1}{N_s}\sum(\text{E}[\boldsymbol{y}'|\boldsymbol{D}] - \boldsymbol{D}')^2}$$

- the corresponding standard deviation of predictions (see Equation 13) as a measure of precision; and

- the posterior predictive coverage based on 90% Bayesian credible intervals between the 5% and 95% quantiles.

These were evaluated for individual models (see Equation 7) and for the BMA pdf-average (see Equation 11). The validation showed that both, the informed zonated model (2a) and the pilot-point model (3), provide the best individual trade-offs in terms of accuracy and precision. However, no single model was clearly superior over the other candidates as can be seen in Table 4. Expect for the homogeneous model which performs worse, the other three candidates show very similar performance on $\boldsymbol{D}'$ when conditioned on $\boldsymbol{D}$. The BMA-average (dominated by the informed zonated model) did not outperform the single models either (cf. Section 4.3.4).

Table 4: Predictive performance of individual models regarding accuracy (RMSE), precision (Standard deviation) and predictive coverage from Schöniger et al. (2015a)

Predictor	RMSE (cm)	Std. (cm)	Pred. Cov. (%)
Homogeneous	0.52	0.06	18
Zonated	0.37	0.20	70
Interpolated	0.33	0.22	74
Geostatistical	0.34	0.23	70

For further details refer to Schöniger et al. (2015a).

4.2 Conducting Bayesian Multi-Model Inference

Expanding the previous work, I address conceptual uncertainty in the modelling task not only from the perspective of BMS/BMA but also by Pseudo-BMS/BMA and Bayesian Stacking, evaluating these multi-modelling frameworks with respect to specified $\mathcal{M}$-settings. For comparability to Schöniger et al. (2015a), I use the same data ($\boldsymbol{D}$ and $\boldsymbol{D}'$) as well as the same numerical samples of prior parameter and prediction distributions. The only exception are numerical samples of the uninformed zonated model which contained errors that distort the model weights of BMS/BMA. I regenerated these samples for this thesis and therefore present different results for BMS/BMA when the model 2b is in the set.

First, I propose a new setting called Quasi-$\mathcal{M}$-closed that allows to pursue process-identification in applied modelling tasks. Second, I define the $\mathcal{M}$-settings for the sandbox case study and explain which Bayesian multi-model frameworks are applied in each setting. Third, I show how the marginalized likelihoods required to apply the frameworks can be obtained from prior predictive distribution samples.

4.2.1 Defining a Quasi-$\mathcal{M}$-closed setting

The three $\mathcal{M}$-settings of Bernardo and Smith (1994), $\mathcal{M}$-closed, $\mathcal{M}$-complete and $\mathcal{M}$-open, are comprehensive. However, in physical sciences or engineering, models are often built that contain all known details about the physical system at the current state of science - so they are considered to be very close to the truth. Yet, knowing that this knowledge is incomplete prohibits to fully assume to be in the $\mathcal{M}$-closed setting. Further, even if all physics of the system were fully represented in a model, there are still issues that come from model computation: Unless analytical solutions are available, numeric discretization schemes produce errors that range from inaccuracies to numerical artifacts. Then, the model output, e.g., the spatial spread of a solute in groundwater, contains numerical errors that might be misinterpreted as features of the real system but are only a result of shortcomings in the applied numerics.

Hence, I want to introduce a fourth setting, denoted Quasi-$\mathcal{M}$-closed, which enables us to pursue the identification of the DGP even if we cannot (yet) fully write down the exact mathematical description for it or have to assume numeric inaccuracy. Thereby, we follow the terminology of Burnham and Anderson (2004) who call the targeted model under these circumstances *quasi-true*. Formally, the setting is $\mathcal{M}$-complete, because the true model is not exactly represented by one of the candidate models. Still, for practical purposes, it is treated as $\mathcal{M}$-closed such that methods for identification like BMS/BMA can be applied. This practical

compromise comes with the restriction that the best model is not the true model - we rather have to interpret the best model from a D_{KL} perspective as having a prior predictive distribution that is very close to the pursued (but yet unavailable) $q(\boldsymbol{y}|M_{\text{true}})$.

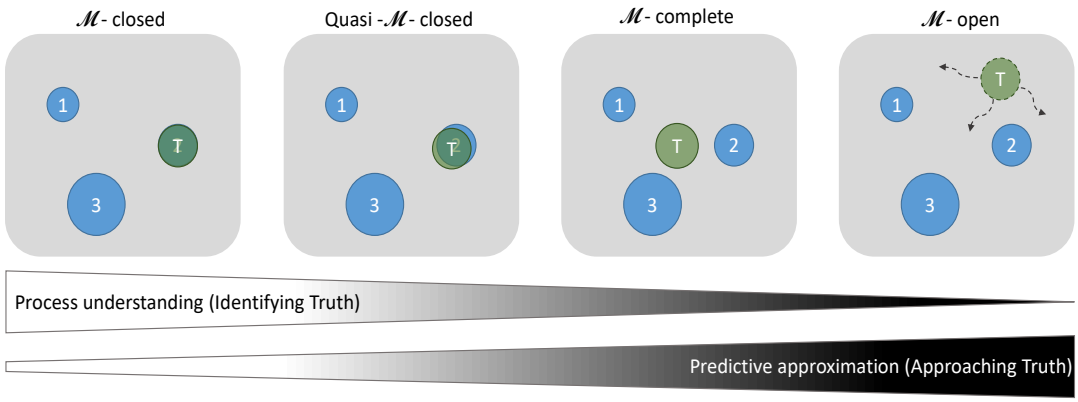


Figure 16: Illustration of the four $\mathcal{M}$-settings as 2D projection: $\mathcal{M}$-closed (left), Quasi-$\mathcal{M}$-closed (center left), $\mathcal{M}$-complete (center right) and $\mathcal{M}$-open (right). The model set comprises three models (blue circles) of different complexity (indicated by the circle size). While in the $\mathcal{M}$-closed, Quasi-$\mathcal{M}$-closed and $\mathcal{M}$-complete setting the true model (green circle with "T") is static in the model space, arrows in the $\mathcal{M}$-open setting depict the true model as "moving target". The primary objective (process-understanding or predictive approximation) in each setting is visualized by the grey scale (bottom).

As can be seen in Figure 2, Quasi-$\mathcal{M}$-closed can be allocated on the black-white scale in the region of light grey, while $\mathcal{M}$-complete rather refers to dark gray. The qualitative differences of the $\mathcal{M}$-settings are summarized in Table 5.

Table 5: Qualitative summary of the four $\mathcal{M}$-settings: $\mathcal{M}$-closed, Quasi-$\mathcal{M}$-closed, $\mathcal{M}$-complete and $\mathcal{M}$-open with respect to the true model.

Model (pdf)...	M-closed	Quasi-M-closed	M-complete	M-open
... can be conceptualized	fully	fully	fully	incompletely
... can be written down	fully	nearly	incompletely	impossibly
... matches actual true model (pdf)	fully	nearly	maybe closely	maybe temporarily

4.2.2 Evaluating Multi-Model Frameworks in Different $\mathcal{M}$-settings

As described in Section 2.1.4, the different $\mathcal{M}$-settings deviate by whether the DGP is assumed to be included or not and by how the list of models in the ensemble relate thereto. In Figure 17 all model parametrizations are illustrated and I aligned them in three different ways to represent the $\mathcal{M}$-closed, the Quasi-$\mathcal{M}$-closed, and the $\mathcal{M}$-complete setting, respectively.

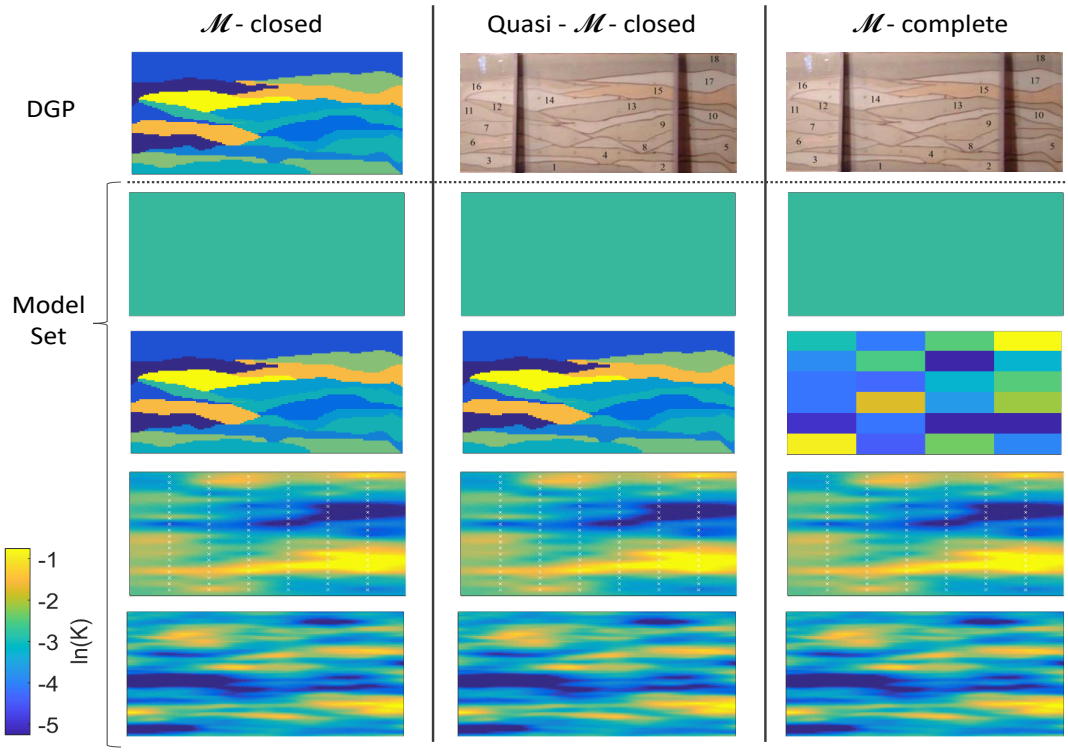


Figure 17: Defined $\mathcal{M}$-settings for the applied modelling example: 1st column) $\mathcal{M}$-closed - DGP: informed zonated model - model set: homogeneous, inf. zonated, pilot-point and geostatistical; 2nd column) Quasi-$\mathcal{M}$-closed - DGP: sandbox - model set: homogeneous, inf. zonated, pilot-point and geostatistical; 3rd column) $\mathcal{M}$-complete - DGP: sandbox - model set: homogeneous, uninf. zonated, pilot-point and geostatistical

In the sandbox example, there is no $\mathcal{M}$-open scenario because we can fully conceptualize the steady-state problem (at least once we accept the continuum assumption to define a hydraulic conductivity). The only relevant hindrance is that we cannot fully resolve what we have conceptualized, e.g., the physical properties in the fringes of the different sand zones.

1. In the $\mathcal{M}$-closed setting (Figure 17, left), identification of the true model (2a) is the logical objective, i.e. consistent model selection. Hence I conduct BMS/BMA to achieve this task, partly reproducing the justifiability analysis of Schöniger et al. (2015a) for model 2a as DGP. I consider the repetitive evaluation of BME on various $\boldsymbol{D}^{syn}$ for obtaining the confusion matrix as extended prior predictive check (see Gabry et al., 2017). To that, I add an extended posterior predictive check: I conduct Pseudo-BMS/BMA in parallel in order to contrast the model rating from BMS/BMA with a method that is not tailored for the identification of the true model. For computation

feasibility, I reduce the 1000 $\boldsymbol{D}^{syn}$ to 100, because the evaluation of Pseudo-BMS/BMA based on prior samples ($2{\cdot}10^5$ and $1{\cdot}10^7$) is computationally more expensive due to the leave-one-out cross validation procedure (see Section 4.2.3). It has to be repeated over all 210 observations of the calibration data throughout all 63 pumping test combinations.

2. Using real-world observations with the same models turns the $\mathcal{M}$-closed into a Quasi-$\mathcal{M}$-closed setting (Figure 17, center). Knowing that the exact true model is not in the set, I apply Pseudo-BMS/BMA to find the single best predictive model. In comparison, I apply BMS/BMA to find the quasi-true model. Besides Pseudo-BMS/BMA for predictive selection, this setting qualifies for the application of Bayesian Stacking to converge to a suitable predictive model combination. Further, due to the limited amount of data points in $\boldsymbol{D}$, I apply the Bayesian Bootstrap (BB) to the two predictive multi-modelling frameworks Pseudo-BMS/BMA and Bayesian Stacking to account for uncertainty in model weighting.

3. Substituting the informed zonated model (2a) by the uninformed zonated model (2b) turns the Quasi-$\mathcal{M}$-closed into an $\mathcal{M}$-complete setting (Figure 17, right). There, only predictive methods for selection or combination are expected to be suitable since no (quasi-)true model can be identified. Hence, I focus on Pseudo-BMS/BMA and Bayesian Stacking, applying the BB to both frameworks. For contrasting their behaviour with (bound to be misleading) consistent model selection, I also apply BMS/BMA in this $\mathcal{M}$-complete setting.

4. In the Quasi-$\mathcal{M}$-closed and $\mathcal{M}$-complete settings, I evaluate all three frameworks with respect to their predictive performance on the validation data. Therefore, I calculate the same measures for accuracy (RMSE), precision (standard deviation) and predictive coverage (90 % Bayesian credible intervals) as done by Schöniger et al. (2015a) for comparability.

4.2.3 Obtaining the Marginalized Likelihoods

Each Bayesian multi-model framework rests on the evaluation of a marginalized likelihood to quantify model performance in its respective realm. To guarantee comparability between the three frameworks defined in Section 2.3, all model scores are evaluated on the same prior samples for each model. In a straight-forward manner, each model's marginal likelihood $p(\boldsymbol{D})$ (or $p(\boldsymbol{D}_{\varnothing})$) is obtained by plain Monte Carlo integration (Equation 22). $p(D_o|\boldsymbol{D}_{\varnothing})$ is then gained by exploiting the following relation:

$$
\begin{aligned}
p(D_o|\boldsymbol{D}_{\varnothing}) &= \int p(D_o|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{D}_{\varnothing})\mathrm{d}\boldsymbol{\Theta} \\
&= \int p(D_o|\boldsymbol{\Theta})\frac{p(\boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\boldsymbol{D}_{\varnothing})}\mathrm{d}\boldsymbol{\Theta} = \frac{\int p(D_o|\boldsymbol{\Theta})p(\boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int p(\boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} \\
&\overset{\text{iid}}{=} \frac{\int p(D_o, \boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int p(\boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{\int p(\boldsymbol{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}}{\int p(\boldsymbol{D}_{\varnothing}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} = \frac{p(\boldsymbol{D})}{p(\boldsymbol{D}_{\varnothing})}
\end{aligned}
\tag{42}
$$

The i.i.d. assumption of measurement error, i.e., conditional independence of D_o and $\boldsymbol{D}_{\varnothing}$ given parameters, also applies here in form of the uncorrelated Gaussian likelihood function. Therefore, the point-wise predictive density $p(D_o|\boldsymbol{D}_{\varnothing})$ can be understood as the ratio between the marginal likelihoods (BMEs) of the whole data set $\boldsymbol{D}$ and LOO data set $\boldsymbol{D}_{\varnothing}$. This elucidates why predictive (non-consistent) model selection does not support the conclusion that a certain model has generated the data (as in consistent model selection) - the likelihood that the model generated "past" data $\boldsymbol{D}_{\varnothing}$ drops out and only the predictive density for the "future" data point D_o remains.

Going pointwise through the whole data set $\boldsymbol{D}$ yields an expression of the expected logarithmic pointwise predictive density for model rating (see Equation 18) by using Equation 42 that writes as:

$$
elpd_{LOO} = \sum_{o=1}^{N_s} \ln \frac{p(\boldsymbol{D})}{p(\boldsymbol{D}_{\varnothing})} = N_s \ln p(\boldsymbol{D}) - \sum_{o=1}^{N_s} \ln p(\boldsymbol{D}_{\varnothing}) \tag{43}
$$

This formulation further allows for an interpretation of $elpd_{LOO}$ from the perspective of explicit information criteria (cf. Section 3), considering the log-BME as decomposed into a sum of a "goodness-of-fit" and a "model complexity" term like, e.g., in the KIC: In consistent model selection, we know that the model complexity has to grow subextensively with increasing amount of data (Bialek et al., 2001) to allow for convergence toward the true model. Hence, the substraction of log-BME values for all but one data point from the full log-BME in equation 43 can be interpreted as their subextensively growing model complexity terms to cancel each other out up to the contribution from the left out data point. What remains is a non-consistent metric without the property to converge towards a presumably true model. Using the $elpd_{LOO}$ for model evaluation only focuses on rating the model in its ability of predicting a single next data point - marginalized over all available data points - but without the need to also cover the past data for explanatory purposes.

The obtained marginalized likelihoods are then turned into model weights for all three Bayesian multi-model frameworks by employing equations 14, 19 and 20. An analytic solution to the marginalized likelihoods and their interrelation can be obtained for Gaussian linear models as shown in Appendix C.

4.3 Results and Discussion

4.3.1 Model Weights in the $\mathcal{M}$-closed Setting

For the $\mathcal{M}$-closed setting, the two compared frameworks BMS/BMA and Pseudo-BMS/BMA yield similar results. Figure 18 shows that both frameworks clearly prefer the zonated model in a selection. This does not come as a surprise because from the perspective of prior or posterior predictive checks, the true model will always yield best predictions.

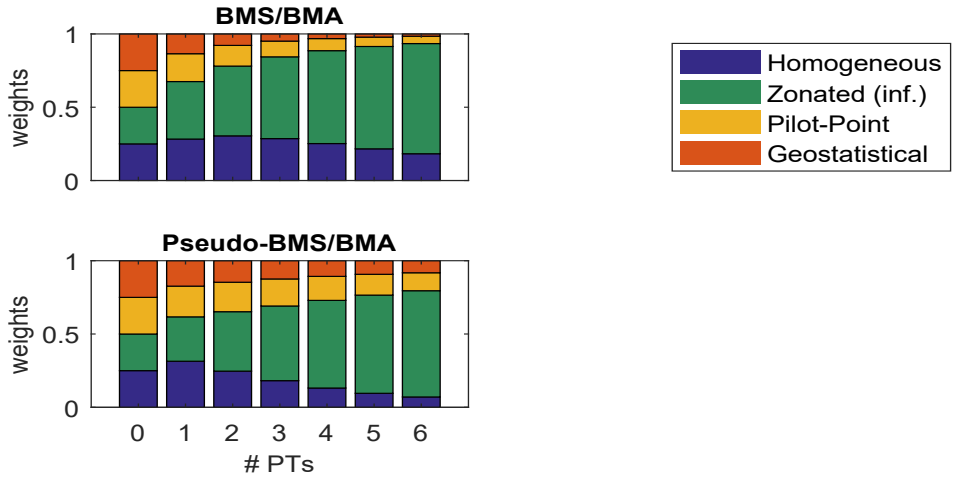


Figure 18: Expected model weights over growing data size (number of included pumping tests, # PTs) for the $\mathcal{M}$-closed setting, i.e., with data generated by the informed zonated model (DGP). Top: BMS/BMA; bottom: Pseudo-BMS/BMA.

Plainly spoken, BMS/BMA simply does its job and converges towards the informed zonated model which is in fact the DGP. By construction, its prior predictive distribution is "closest" (identical) to the true data distribution. Due to the limited amount of (calibration) data, the zonated model reaches a maximum model weight of 75% when all 210 data points are included in the inference. Despite using only 100 realizations of $\boldsymbol{D}^{syn}$, this confirms the result of the justifiability analysis of Schöniger et al. (2015a). This confirmation verifies also the reliability of the Pseudo-BMS/BMA results. In the additionally evaluated Pseudo-BMS/BMA, the zonated model also receives the highest weight, as shown on the bottom of Figure

18. Obviously, its posterior predictive distribution is closest to the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$ in the D_{KL} sense, too.

However, there are several differences to be noticed between the two frameworks: At each data stage, the obtained model weights deviate between the two frameworks which becomes more and more distinct over growing data size. In BMS/BMA, the weights of more complex model alternatives (3 and 4) immediately decline as data size increases and with all 6 pumping tests included, only the simpler homogeneous model remains as weak but still relevant alternative to the informed zonated model. In Pseudo-BMS/BMA it is vice versa - after a slight peak for considering only one pumping test, the weight of the simpler model diminishes. Yet, the framework yields still significant weights for the more complex alternatives due to their high predictive capability. This confirms both the conservative nature of consistent model selection methods and the tendency of non-consistent methods towards more complex models (see Section 3).

Yet, the most fundamental difference is that only the BMS/BMA weights can be interpreted as probabilities in terms of the zonated model being the true model due to the consistency property of the multi-model framework. Although the Pseudo-BMS/BMA weights indicate largest (posterior) predictive capability of the zonated model, it cannot be interpreted as true model. The underlying assumption behind Pseudo-BMS/BMA - being outside of $\mathcal{M}$-closed - prohibits this interpretation even if a model received a model weight of exactly one.

4.3.2 Model Weights in the Quasi-$\mathcal{M}$-closed Setting

In the Quasi-$\mathcal{M}$-closed setting, the focus shifts stronger to predictive Bayesian multi-model frameworks. This requires to account for limited observations with the Bayesian Bootstrap. For the two predictive methods, Pseudo-BMS/BMA and Bayesian Stacking, the bootstrapped results are decisive for the comparison between the three methods. The frameworks yield vastly different results, as it can clearly be seen in Figure 19.

BMS/BMA assumes to be employed in an $\mathcal{M}$-closed setting and (incorrectly) identifies the zonated model (2a) as DGP. Knowing, that this model is not the exact representation of the actual DGP, we can still take it as the most plausible quasi-true model. This preference for the zonated model is similar to the result in Section 4.3.1, yet the model alternatives are rated differently. In the Quasi-$\mathcal{M}$-closed setting, the simple homogeneous model is essentially discarded despite the conservative tendency of BMS/BMA in model selection. Instead, the more complex pilot-point model (3) remains as plausible candidate and even the most

complex geostatistical model (4) receives higher probability to be the (quasi-)true model than the simplest one. This shows that, despite the zonated model appearing as presumably close resemblance of the true physical system, BMS/BMA struggles to identify it as DGP as clear as in an actual $\mathcal{M}$-closed setting.

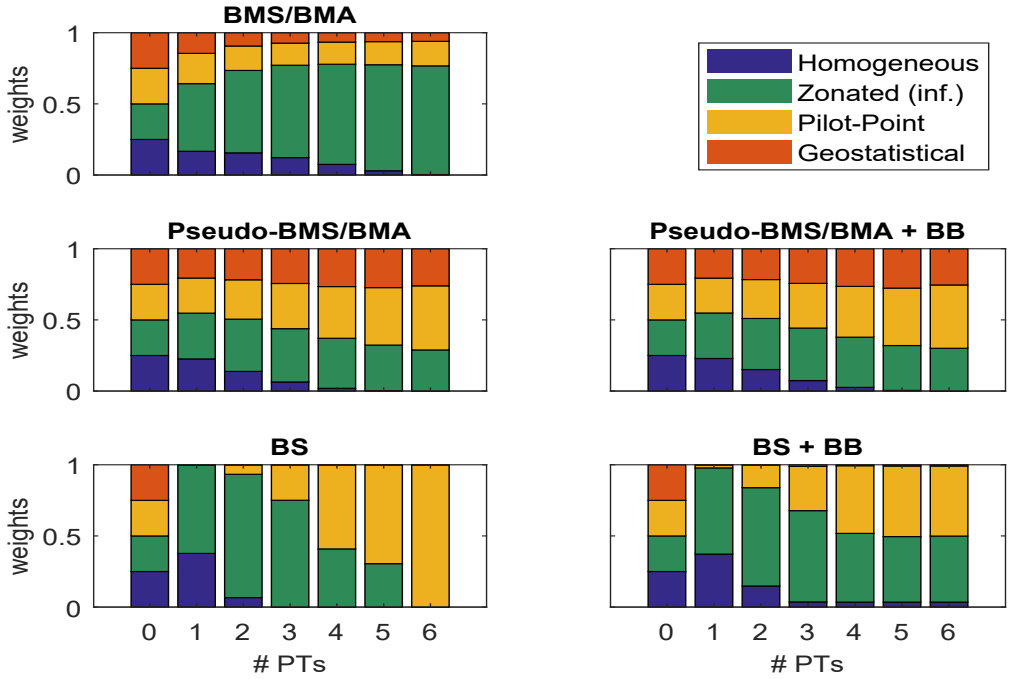


Figure 19: Expected model weights over growing data size (number of included pumping tests, # PTs) for the Quasi-$\mathcal{M}$-closed setting, i.e., with data observed from the sandbox aquifer. Top row: BMS/BMA; center row: Pseudo-BMS/BMA without (left) and with (right) Bayesian Bootstrapping (BB); bottom row: Bayesian Stacking (BS) without (left) and with (right) BB.

Pseudo-BMS/BMA yields a very different result in Quasi-$\mathcal{M}$-closed when directly compared to BMS/BMA and also in comparison to the result of Pseudo-BMS/BMA in $\mathcal{M}$-closed (cf. Section 4.3.2). The tendency toward models of higher complexity is obvious and the simplest model (1) never stands a chance to be preferred. In non-consistent selection, the model that promises largest predictive capability while having a complexity that is still supported by the current amount data is preferred. Hence, even though a human expert might think that the zonated model is the closest resemblance of the physical sandbox, Pseudo-BMS/BMA finds this trade-off to be fulfilled best by the pilot-point model (3) over growing data. Over increasing amount of data, the tendency toward more complex models is clearly visible. While for two pumping tests, the zonated model is rated best, yet without strong advance, it loses ground in favor of the more complex alternatives - especially the pilot-point model. The most complex geostatistical model does

not gain weight but (contrarily to the BMS/BMA weights and to the results from $\mathcal{M}$-closed) it also does not loose any. The Bayesian Bootstrap confirms the model weights as being robust over various instances of $\boldsymbol{D}$ from the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$. Hence, the most likely bootstrapped weights are almost equal to the original Pseudo-BMS/BMA weights.

Bayesian Stacking for predictive model combination behaves completely differently from either selection framework. The conjoint inference of model weights via rating the combined ensemble $\mathcal{K}$ is more challenging than rating each model individually and afterwards evaluating the weights as in Pseudo-BMS/BMA: The maximization problem in Equation 20 has several local maxima as solutions for different instances of $\boldsymbol{D}$ and Bayesian Stacking without bootstrapping yields only one. Hence, the BB shows a much larger effect, yielding the global maximum over the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$. The converged $\boldsymbol{w}^{\text{BB}}$ are, again, more robust and allow for a more reliable interpretation. The stabilized model weights distribution over four to six pumping tests supports the assumption of convergence. At each data stage, Bayesian Stacking seeks to find the combined average of models that covers the data best. While for a single pumping test this seems to be accomplished by a combination of the homogeneous and the zonated model, more complex models receive higher weight in the combination with growing data. Over two and three pumping tests, Bayesian Stacking moves toward a stable combination of essentially the pilot-point and the zonated model (with minor contribution of the homogeneous one). It appears to combine the preferred models of both BMS/BMA and Pseudo-BMS/BMA. While the BB adjustment is not as strong for small data sizes it has more impact for growing data when it is more difficult to find a global maximum. For four and more pumping tests, the stable model weights allow for averaging in the sense of real model combination. In an illustration like Figure 2, the true DGP can conceptually be located roughly between the pilot-point and zonated model. Physically, this can be interpreted as accounting for the fringes of the sand layers: The zonated model shows too stark contrasts between neighbouring zones and the PP shows a too smooth transition. The reality is probably somewhere in between, and the Bayesian Stacking weights reflect this.

4.3.3 Model Weights in the $\mathcal{M}$-complete Setting

For the $\mathcal{M}$-complete setting, the focus is only on predictive Bayesian multi-model frameworks supported by Bayesian Bootstrapping. From the underlying theory, BMS/BMA is inadequate to be applied in this setting. Its results are only presented to contrast it with predictive methods, taking into account that it is often applied nonetheless in $\mathcal{M}$-complete settings like a "panacea" (Clyde and Iversen, 2013). Accordingly, the results of the three frameworks deviate strongly, as shown

in Figure 20.

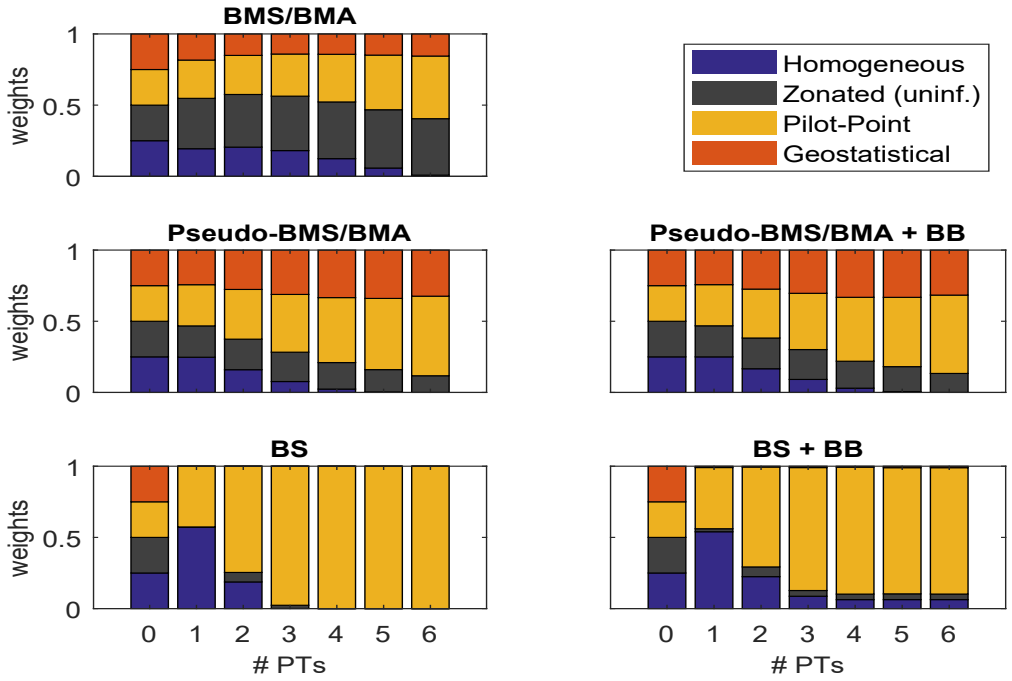


Figure 20: Expected model weights over growing data size (number of included pumping tests, # PTs) for the $\mathcal{M}$-complete setting, i.e., with data observed from the sandbox aquifer. Top row: BMS/BMA; center row: Pseudo-BMS/BMA without (left) and with (right) Bayesian Bootstrapping (BB); bottom row: Bayesian Stacking (BS) without (left) and with (right) BB.

Due to its consistent nature, BMS/BMA seeks to identify one of the models as (quasi-)true but it is not able to clearly prefer only one over the others. Up to the inclusion of three pumping tests, BMS/BMA shows a preference for the uninformed zonated model (2b), mostly due to its parsimony. However, this preference for a single candidate is by far not as clear as in the $\mathcal{M}$-closed or Quasi-$\mathcal{M}$-closed settings. For four and more pumping tests, the pilot-point model (3) receives growing weight while the homogeneous model (1) loses ground and the geostatistical model (4) remains nearly constant. In the $\mathcal{M}$-complete setting, BMS/BMA only yields an indecisive selection in the investigated range of data availability. Yet, again, this shall not be confused with converging to an average in terms of model combination, because BMS/BMA does not have the property to do so.

From its underlying theory, Pseudo-BMS/BMA is supposed to be a suitable choice in an $\mathcal{M}$-complete setting and yields plausible results in contrast to BMS/BMA. The framework behaves similarly to the performance in Section 4.3.2), yet it is even stronger in its preference for the pilot-point model (3) - simply because (2b)

is not a good competitor. The principal preference for complex models is depicted by the geostatistical model (4) obtaining the second highest weight when at least two pumping tests are considered, which makes it clearly the second choice in a selection for predictive purposes. In the same line of thought, the simple homogeneous model (1) diminishes in weight over the successive inclusion of up to three pumping tests and the uninformed zonated model (2b) also has declining model weight because it comes with comparably large complexity given insufficient fit to data. Apparently, in the inappropriate zonation structure, the available complexity is spent on comparably useless model degrees of freedom. Again, the Bayesian Bootstrap confirms the model weights from Pseudo-BMS/BMA (cf. Section 4.3.2).

In Bayesian Stacking, the pilot-point model also receives highest weight rather quickly, not as preferred single model but rather as dominating fraction in a model combination. The BB shows similar behaviour as in the Quasi-$\mathcal{M}$-closed setting, accounting for the difficulty of joint weighting under yet growing but still limited data size. Hence, Bayesian Stacking with Bayesian Bootstrapping also converges to stable fractions of models for four to six pumping tests. In contrast to Quasi-$\mathcal{M}$-closed, the Bayesian Stacking model average in the $\mathcal{M}$-complete setting relies nearly entirely on the pilot-point model (3). A small fraction in the combination is roughly equally covered by the homogeneous and uninformed zonated model. The geostatistical model is not part of the combination - due to its large complexity, forecasts from this model only come with low predictive density which makes it insignificant in the weighted average that is optimized for high predictive density. In the illustration like in Figure 16, the predictive distribution is very close to the pilot-point model, slightly pulled toward the simpler models. This can again physically be interpreted as compensating for the too smooth fields from the pilot-point model. Yet, in the $\mathcal{M}$-complete setting, the compensation does not honour the geometric boundaries of the different zones as in Quasi-$\mathcal{M}$-closed and therefore the simpler models contribute only marginally.

4.3.4 Validation in Quasi-$\mathcal{M}$-closed and $\mathcal{M}$-complete settings

The validation with further observations $\boldsymbol{D}'$ was performed for the Quasi-$\mathcal{M}$-closed and $\mathcal{M}$-complete settings. The only clear message thereof is: All multi-model frameworks yield similar results for the chosen application example.

No framework shows its superior predictive capability, neither in the Quasi-$\mathcal{M}$-closed nor the $\mathcal{M}$-complete case. A possible explanation therefore concerns the used data: two of the four additional pumping tests that provided the validation data (140 data points) were performed in parts of the sandbox, about which no information was contained in the calibration data as already mentioned by Schöni-

ger et al. (2015a). This questions whether there was already enough calibration data used with the six pumping tests to allow for sufficient inference of the model with high parametric complexity. Yet, one might argue that a main purpose of modelling is to provide predictions under conditions never experienced before. Therefore, calibration data should be able to inform and constrain the employed model(s) in a way that regardless from when and where validation data is observed, it should be matched by the model(s) to some degree. Hence, if the individual members of a model set struggle with extrapolating to new situations, it cannot be expected that a multi-model framework per se yields better results. In the first place, the model set members require adjustment and improvement.

Mainly, the results can yet be explained with the apparent similarity of the individual models regarding their individual predictive performance. Looking at Table 4, only the homogeneous model shows significantly higher RMSE, lower standard deviation and lower predictive coverage. The other three models (2a, 3 and 4) yield very similar results regarding accuracy, precision, and coverage.

Hence, in the Quasi-$\mathcal{M}$-closed setting, no weighted average contains a significant contribution of the homogeneous model. Predominantly, the averages are based on the informed zonated (2a) and pilot-point model (3) - only in Pseudo-BMS/BMA also the geostatistical model (4) obtains significant weight. Since all of them provide similar prediction metrics, their weighted averages do as well as presented in Table 6. Although the model weights honour the objective (DGP identification vs. best individual predictive model vs. best predictive combination) of each respective Bayesian multi-model framework in distinctly different weightings, their predictive results are similar.

Table 6: Predictive performance of model averages from Bayesian model frameworks (and the informed zonated model) in the Quasi-$\mathcal{M}$-closed setting regarding accuracy (RMSE), precision (Standard deviation) and predictive coverage. (Note, that the results for BMS/BMA differ from the ones reported in Schöniger et al. (2015a) where, presumably, typos made it into the printed version of the article.)

Predictor	RMSE (cm)	Std. (cm)	Pred. Cov. (%)
zonated (inf.)	0.37	0.20	70
BMS/BMA	0.36	0.21	73
Pseudo-BMS/BMA	0.34	0.22	73
Bayesian Stacking	0.35	0.22	72

The same holds for the $\mathcal{M}$-complete setting with the uninformed zonated model (2b) as shown in Table 7. Except for the lower predictive coverage of 64%,

this model shows similar individual performance to the more complex models (3) and (4) regarding RMSE and standard deviation. The averages from the three frameworks show, again, only negligible differences: the RMSEs of 0.34 cm and the standard deviations of 0.22 cm are identical in all frameworks and the predictive coverages are very similar - although these results were achieved by vastly different weightings of the models within each framework. The highest weighted model in all averages is the pilot-point model, but the uninformed zonated model is weighted comparably strongly in the non-decisive BMS/BMA and the the geostatistical model obtains significant second largest weight in Pseudo-BMS/BMA. Hence, all rating results are in full compliance with the respective framework objective. Yet, due to their similar individual performances, this has negligible effect on the average performance metrics of the respective frameworks.

Table 7: Predictive performance of model averages from Bayesian model frameworks (and the uninformed zonated model) in the $\mathcal{M}$-complete setting regarding accuracy (RMSE), precision (Standard deviation) and predictive coverage

Predictor	RMSE (cm)	std (cm)	Pred. Cov. (%)
zonated (uninf.)	0.35	0.19	64
BMS/BMA	0.34	0.22	71
Pseudo-BMS/BMA	0.34	0.22	70
Bayesian Stacking	0.34	0.22	72

Although the validation results of this example do not clearly depict the respective strengths of each method, they neither support that it is irrelevant which multi-model framework is employed. They rather emphasize the necessity to correctly interpret the meaning of the model weights of each framework that underpin these validation results:

- Bayesian Stacking seeks to compensate for shortcomings of individual models regarding posterior predictive density by shares of model alternatives and converges to model fractions that are optimal in this respect. It therefore yields a best possible compromise between models, balancing the model set by averaging for comprehensiveness.
- Both BMS/BMA and Pseudo-BMS/BMA are ultimately selection frameworks that indicate to use only the model with highest weight. Yet, to avoid misleading results before a completed selection (model weight of 1 for one model), averaging assures reliability. Naturally, averages are not better than an individual best model, but they account for remaining conceptual uncertainty.

4.4 Summary and Conclusion

Only by accounting for the $\mathcal{M}$-setting of the modelling task, Bayesian multi-model frameworks can be properly applied and one can exploit their full potential:

- In the $\mathcal{M}$-closed setting, BMS/BMA allows to identify the true model by model probabilities and thereby yields highest predictive capability for unseen data. Pseudo-BMS/BMA might prefer the true model, too, but it does not allow to identify it as such.
- In the Quasi-$\mathcal{M}$-closed setting, BMS/BMA might converge to the closest resemblance of the truth but one cannot expect to obtain best future predictions by it. Pseudo-BMS/BMA might prefer another model as best in predictive model rating. Alternatively, Bayesian Stacking might yield a model combination optimized for predictions that depicts a compromise between suitable models.
- In the $\mathcal{M}$-complete setting, BMS/BMA will be misleading, let alone that it is able to converge to one candidate at all. In this setting, only predictive methods for selection and combination are suitable and the modeller has to decide whether an individual model or a weighted average shall be employed for predictions.

Hence, before any Bayesian multi-model framework is applied in a practical modelling task, it has to be clearly defined in which setting we allocate the modelling task and which goal we pursue there:

- If the purpose is to understand the DGP, we follow the Quasi-$\mathcal{M}$-closed perspective whenever possible. There, we apply BMS/BMA, knowing that a direct interpretation of the model weights as probabilities is yet impossible but seeking to match the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$ with the prior predictive distribution $p(\boldsymbol{y})$ of one of our models in a D_{KL}-sense. The best model identified by this procedure can then be the basis for a next stage of model refinement where we improve this model in a way we think that it resembles the truth even better (here, by a refined parametrization for the fringes in the informed zonated model). Then, having a new set of alternatives, we re-iterate BMS/BMA, and so on. Ultimately, we want to move our modelling task to a full $\mathcal{M}$-closed scenario where our model weights actually resemble model probabilities.
- If the the assumption of having a (quasi-)true model in the set is unreasonable, we follow the $\mathcal{M}$-complete (or even $\mathcal{M}$-open) perspective. There, to

achieve maximum predictive power, we apply Pseudo-BMS/BMA or Bayesian Stacking to approximate the data distribution $q(\boldsymbol{y}|M_{\text{true}})$ with the (individual or common) posterior predictive distribution. The single best model that comes out of Pseudo-BMS/BMA can then be the basis for adding complexity as the available data size increases to approach the unknown DGP even more closely. In the $\mathcal{M}$-complete setting, Bayesian Stacking is supposed to yield a stable model combination that ensures predictive reliability. In an $\mathcal{M}$-open setting, it cannot be expected that Bayesian Stacking converges to stable weights given that the true model is a "moving target". Hence, the continuous improvement of models or enlargement of the model set is indispensable to increase predictive capability.

In case that a clear allocation of the modelling task in a certain $\mathcal{M}$-setting between the extremes, $\mathcal{M}$-closed and $\mathcal{M}$-open, is impossible (see Figure 16), switching between the two perspectives, Quasi-$\mathcal{M}$-closed and $\mathcal{M}$-complete, might be a reasonable practical approach. Yet, it is then crucial to only interpret the model weights accordingly in order to properly address conceptual uncertainty.

Whenever model weights are based on some inferred quantities, it is necessary to account for associated inferential uncertainty, e.g., via the Bayesian Bootstrap. For predictive multi-model frameworks like Pseudo-BMS/BMA or Bayesian Stacking, the available data as proxy for unseen data is typically the source of uncertainty in model weights of highest priority. In process-identification like via BMS/BMA, prioritizing might show that the robustness of model weights suffers mostly from measurement noise of the available data and model inputs, or vagueness in boundary conditions (see Schöniger et al., 2015b). This has to be addressed separately.

Although model validation might provide similar results for all evaluated multi-model frameworks for the currently available data (as in this example), the averages have to be scrutinized with respect to the contributions of each model in the set. It can be expected that significant deviances in predictive performance in terms of accuracy, precision, coverage, etc. arise when more data is included. Sooner or later, the multi-model average will converge to the single best model in both selection frameworks. Contrarily, the stacking result as superposition of predictive pdfs will provide predictive reliability by making use of several model set members in the average. Nonetheless, deficiencies of the models in matching the true data distribution mark an upper bound for (prior or posterior) predictive power in applied modelling. This can be overcome by continuous model set improvement where each iteration has to be rated in Bayesian multi-model frameworks.

This holds even more for models that deviate strongly in terms of their concep-

tuality. In case, e.g., different model types are employed and not just white-box models models with different levels of detail (as in this example), their type-specific limitations and potentials oblige to perform a thorough analysis of conceptual uncertainty in order to exploit their full potential regarding explanatory or predictive power. As demonstrated throughout this chapter, this can only be achieved by properly applying a Bayesian multi-model framework with respect to the underlying model setting.

5 Guiding toward Task-Specific Multi-Model Use

Fully understanding model rating scores (see Chapter 3) and being aware of how to apply and interpret built-on multi-model frameworks in different model settings (see Chapter 4) should enable us to properly employ multiple models. However, it is hardly possible to, first, keep all the relevant details in mind and therefore, second, to reliably find the most suitable multi-model method for any given modelling task.

I introduce a guideline in Section 5.4 that helps to find the suitable framework based on comparably simple questions about the modelling task. These questions concern the goal of the modelling task and the available model set and have to be answered by the modeller. Before, I will disentangle the terminology of model combination in multi-model use in Section 5.1. Since Bayesian Stacking only allows for model combination to increase predictive power, I show how alternative fully developed models in a set can be combined and rated for the sake of process-identification in a consistent manner in Section 5.2. To clarify interpretability of different multi-model averages, I elaborate the distinction between averaging of model outputs and predictive distributions in Section 5.3. I close the chapter with a summary, discussion and conclusion in Section 5.5.

5.1 Disentangling Model Combination Terminology

Many methods are found under the umbrella of *model combination* and many of them use similar terminology. Often, they refer to "stacking" as combination scheme, but mean different things:

- Bayesian Stacking, that I presented and employed in previous chapters, means *stacking predictive distributions*. Again, this does not change the individual predictive distributions of the specified models but builds a weighted average as superposed convex hull. Therefore, the conjoint optimization of model weights in Bayesian Stacking in Section 2.3.3 is based on stacked pointwise predictive densities from individual models instead of the actual model output values.

- Alternatively, stacking can refer to *adding weighted outputs* of fully specified models. The distribution of such a weighted forecast average obtains an own pdf rather than being a linear (convex) mixture of individual pdfs. For this, the weights are obtained differently than the conjoint optimization from Bayesian Stacking (see Section 5.2).

Contrarily, other notions of model combination exist: We might assume that our candidates in the model set allow to be combined in a "blended" model that contains elements from several candidates in order to give a better resemblance of the DGP. In hydrology that might be, e.g., exchanging the parametrization of evapo-transpiration parametrizations or outflow routing-functions between different model candidates (e.g. Clark et al., 2015). The predictive distribution of such a "blended" model might then be closer to the $q(\boldsymbol{y}|M_{\text{true}})$. Yet, unlike the other introduced methods, this approach does not refer to a multi-model framework in the sense of combining individual fully specified models but to build new alternatives to populate the $\mathcal{M}$-space. Therefore, it is not further discussed here.

The same holds for so-called "hybrid modelling", i.e., combination approaches where different model types are used to compensate shortcomings of one model type by a different model type. Examples are data-driven (black-box) approaches that correct numerical errors during the solution of differential equations (see Ray and Hesthaven, 2018) in mechanistic (white-box) models or that are used to account for model parts (in grey or white box models) that cannot be parametrized otherwise (e.g., Mekonnen et al., 2015).

5.2 Bayesian Model Combination for Process Identification

Apart from the discussed circumstances under which predictive model selection (via Pseudo-BMS/BMA) or combination (via Bayesian Stacking) is preferential in multi-model usage (cf. Chapter 4), it remains a fundamental motivation in science and engineering to understand the DGP - and sooner or later to identify the true model. For a single distinct model, this is achieved by BMS/BMA. Often, this "search for the truth" is also the intention when modellers refer to combination of fully specified models for the same purpose - which might be falsely assumed by modellers to be provided by BMA.

I want to specify what is exactly meant by an illustrative example from Höge et al. (2019): "For an observed decline in concentration of a substance, two experts might provide two plausible hypotheses. The first expert hypothesizes the concentration decrease results from only microbial consumption (M1) and the second expert claims that solely abiotic reactions cause the decline (M2). Each expert comes up with a model that contains the mathematical formulation of their respective process. If BMA was applied to both models, it would prefer one over the other, and would do so increasingly clearly with more included data from the decline - BMA assumes only one of the two models can be true and tries to identify

it. BMA would not settle with the weights of the two processes like 75 % biotic and 25 % abiotic under growing data size, even if this ratio represented what actually happened in reality. "

One might think that this ratio can be found by Bayesian Stacking. However, Bayesian Stacking would not search for the correct ratio on the process level, but for the optimal shares of the two individual predictive distributions from the two distinct hypotheses to superpose them. Hence, our desired kind of model combination of fully developed models is on the process level and therefore different from Bayesian Stacking.

What we want is to identify the representation of the process as "stacked" models (Minka, 2002) on process level, i.e. as superposition of individual models' outputs and not of their pdfs (see Section 5.1). A combined model in this sense is a weighted average of the mathematical model equations. The prediction of this combined model is the accordingly averaged individual model outputs. The right combination of models is supposed to represent the DGP and our goal is therefore to identify it - like the superposition of 75 % biotic model M1 and 25 % abiotic model M2 in the above example.

Bayesian Combined Model Selection and Averaging

Such an approach was originally prosed by Monteith et al. (2011) as *Bayesian model combination*. It rates combinations of models in line with consistent Bayesian model selection to fulfil the identification-purpose. The term *Bayesian model combination* is concise but might cause confusion although it simply refers to the methodology of BMS/BMA applied to combined models CM_k instead of individual models M_m. The combined models are defined as:

$$CM_k = w_1 M_1 + w_2 M_2 + ... + w_{N_M} M_{N_M} \tag{44}$$

The models $M_1, ..., M_{N_M}$ are the fully specified models in the model set $\boldsymbol{M}$. The weights w_m are not found by the method, but proposed by the modeller or provided otherwise - in Monteith et al. (2011), the weights stem from an assigned distribution. Then, the same equations as for BMS/BMA (see Section 2.3.1) are applied to the defined CM_k. The computational effort is only slightly larger than applying BMS/BMA to the individual models because both frameworks require full marginalization over each individual model's prior parameter distribution.

The underlying theory and its consistent behaviour to identify the true model then holds for the rated combined models. Therefore, in a straight-forward manner, this

method is called *Bayesian Combined Model Selection/Averaging* (BCMS/BCMA) (Höge et al., 2019). Note, that the weights in Equation 44 do not express conceptual uncertainty between the individual models as in BMS/BMA. In BCMS/BCMA, conceptual uncertainty refers to selecting one combined model as (quasi-)true. Therefore, the model weights (probabilities) for all combined models CM that come from applying the equations in Section 2.3.1 express conceptual uncertainty in this context.

While BMS/BMA converges to the individual model with a prior predictive distribution that is closest to the true data distribution $q(\boldsymbol{y}|M_{\text{true}})$ (Minka, 2002; Monteith et al., 2011), BCMS/BCMA converges to the optimal combined model CM_{opt} with a $p(\boldsymbol{y}|CM_{\text{opt}})$ that is closest. For an application example of BCMS/BCMA to classification problems refer to Kim and Ghahramani (2012). In hydrosystem modelling, an approach that works similarly can be found in Ajami et al. (2007).

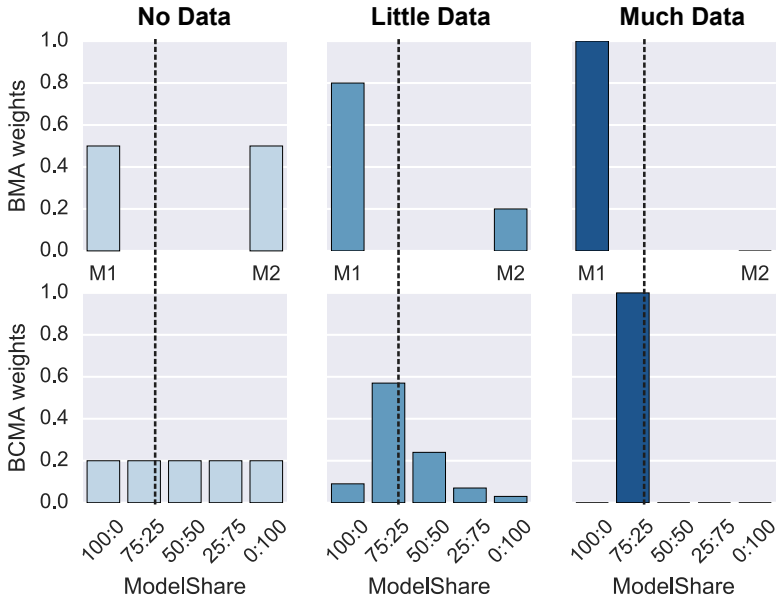


Figure 21: Identification of the model closest to the truth in the set with BMA/BMS (upper half) vs. identification of the most plausible combined model with BCMA/BCMS (lower half) under growing data size (from left to right). The true model (vertical dashed line) is situated between the two model candidates, M1 and M2. In BMA/BMS, weights are assigned to the two distinct models; in BCMA/BCMS, weights are assigned to combinations of both models (M1:M2, ratios in percent). BMA/BMS converges towards one model candidate, BCMA/BCMS converges towards a specific model combination (from Höge et al., 2019).

As discussed by Höge et al. (2019), for the illustrative example above hence follows: "The difference between BMA/BMS and BCMA/BCMS becomes apparent when looking at the change in model weightings under growing data size. This is

illustrated for a simple two-model-setup in Figure 21, like the "biotic vs. abiotic decay" example from above. Without any data, indifferent uniform prior model weights are assigned to the two individual models M1 (microbial) and M2 (abiotic) in BMA/BMS, i.e. 50% each, and for all a-priori specified combinations of the two models in BCMA/BCMS, i.e. exemplary 5 combinations with 20% each. Referring to the conceptual example above, each combined model consists of both the biotic and abiotic reaction terms but by different fractions which resembles that the concentration decrease is caused by both, e.g., 25% microbial and 75% pure chemical decay. Once the individual or combined models face a small amount of data, the model set member closest to the data gains strongest in weight and others gain less or lose model weight. These weights represent the uncertainty in BMA or BCMA of an individual or combined model, respectively, to represent the truth given the current data. Under more and more additional informative data, the weighting converges fully to the one most plausible member in the set: BMA turns into BMS for an individual model and BCMA turns into BCMS for a combined model. In a situation as visualized in Figure 21, where the truth lays somewhere between M1 and M2, BMA/BMS will tend towards the one single model in the set that appears to be most likely to have generated the data - either the biotic or abiotic model but not a mixture. Identifying a truth consisting of combined models will only be pursued by BCMA/BCMS, where the combinations are a-priori defined by the modeler and offered as candidates."

5.3 Averaging of Model Outputs vs. Predictive Distributions

Averaging of models for combining them is always a promising approach if the individual models in the set are assumed to "encircled" the true model (see Sections 2.1.4 and 4.2.1). Yet, it is thereby indispensable to distinguish between averaging of model outputs and averaging of models' predictive distributions:

- Averaging of predictive distributions, as it is done for model combination in Bayesian Stacking or as preliminary result previous to model selection in BMS/BMA or Pseudo-BMS/BMA, yields a mixture of pdfs. This envelope of multiple pdfs is often wide and multi-modal which expresses the associated conceptual uncertainty. Thereby, averaging does not refer to weighted averages of point-estimates from specific (e.g., maximum likelihood) parameters. The statistical moments of the mixture, like mean or variance, should therefore also only be interpreted statistically and not in a physical sense.

- Averaging of model outputs, as it is done to construct combined models according to Equation 44, yields predictions in which deficits of the individual

models are ideally mutually compensated. The pdf of a CM output is supposed to be more focused than the ones of the individual models, providing, e.g., a lower variance than the lowest individual variance (Bates and Granger, 1969). The averaged outputs allow to be interpreted from a physical perspective, in particular in a special case: if each individual model complies with conservation laws and convex (simplex) weights are used, this holds also for the combined model output.

When averaging model outputs, we further have to distinguish between methods with identification-driven or prediction-driven purpose. BCMS/BCMA seeks to identify a combined model as DGP (or rather its $q(\boldsymbol{y}|M_{\text{true}})$) by its prior predictive pdf over averaged model outputs $p(\boldsymbol{y}|CM_{\text{opt}})$ and therefore relies on being provided with a set of combined models.

An overview about methods of model combination that are based on predictions that were originally developed for deterministic model forecasts, is given by Höge et al. (2019): "A variety of model combination weighting schemes on forecast-level has been proposed in order to produce a most accurate combined model for an unknown truth, e.g., equal weights averaging (EWA) or Bates-Granger model averaging (BGA) with weights based on the forecast variance. Most averaging rules rely on "simplex weights" (positive weights that sum to one). Granger and Ramanathan (1984) have proposed to move away from this constraint in order to improve predictive performance of the combined estimator via bias correction (Granger-Ramanathan averaging, GRA). Further, so-called ensemble methods like bagging or boosting exist (Kim and Ghahramani, 2012, and references therein), where combination implies mainly setting up model ensembles and applying distinct model training schemes in a way that the ensemble members mutually counteract bias or variance of forecasts. A comparison of combination approaches for hydrological applications, yet without consideration of their philosophical differences, has been performed by, e.g., Diks and Vrugt (2010)."

The numerous nuances of model combination as well as the previously discussed differences in model selection and averaging often lead to confusion among modellers that want use multi-model frameworks. To support proper choice and utilization, I propose a guide that helps to find the suitable method in the following.

The following Section 5.4 has been published by Höge et al. (2019) and I reuse the text and the corresponding figure. Considering my co-authors, "I" is substituted by "we".

5.4 Guideline to Identify the Best-Suited Multi-Model Approach

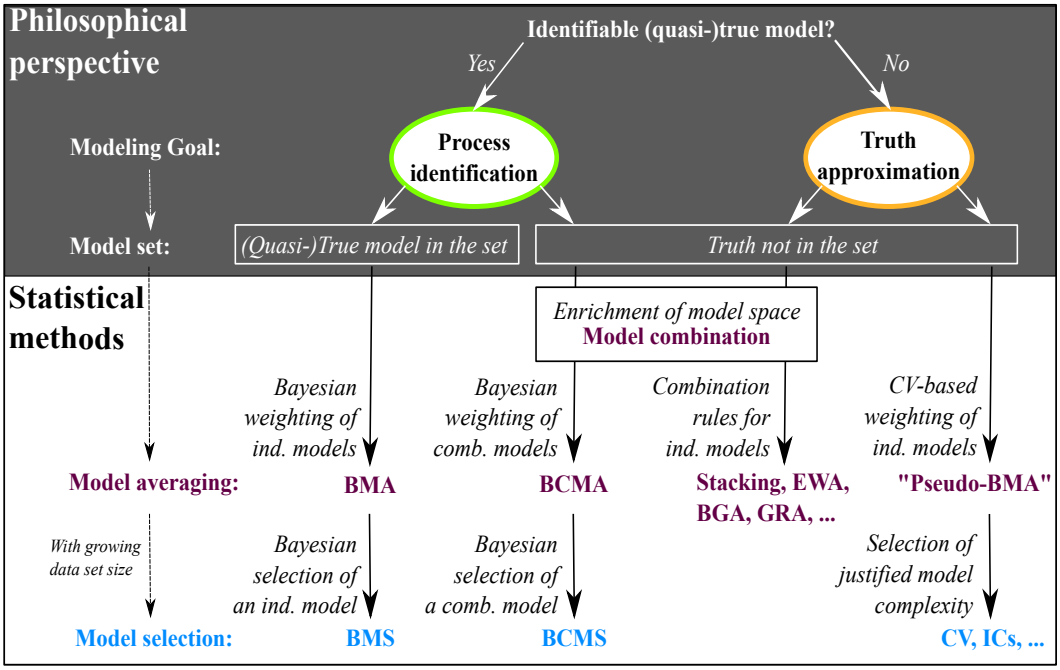


Figure 22: The hydrologist's towel in outer model space: flowchart to find the most suitable statistical approach for multi-modeling. Dashed arrows in the leftmost column guide along the principal steps from the philosophy to the methodical stages and their convergence with growing data. Arrows in the other four columns contain multi-modeling paths from averaging to selection that naturally emerge from specific scenarios of modeling goal and model set assumptions (from Höge et al., 2019).

To not throw in the towel, we propose a scheme for finding the most suitable multi-modeling approach for any modeling task at hand. Starting from the philosophical perspective enables us to find a clear way through the model selection and averaging tree as depicted in Figure 22. BMS is our way to go if we are interested in process identification and the process is represented as identifiable target model in our set of models. Depending on the data availability, this path will pass through BMA which enables us to handle the uncertainty between all plausible hypotheses probabilistically until one reveals itself as best representation of the truth. However, if we think that this target model is not a single model but rather a combined model which is situated somewhere between these candidates in the set, we can propose and select a corresponding combined model using BCMA and ultimately BCMS. In Figure 22, we see that all, BMS, BMA, BCMS and BCMA, are aiming at model identification. By applying BCMA/BCMS, we avoid the uneasy assumption of having a (quasi-)true model in the set immediately, but still we

aim to identify a best combined model that represents the truth.

If we have to deny the initial question about whether there is an identifiable true model or at least a quasi-true model, we can only seek to approximate this truth to obtain plausible predictions. Then, our options are either to enrich model space via classic combination approaches (EWA, BGA, GRA,...), or to select a single model that promises the best while parsimonious approximation. Actual model combination approaches as in the former option estimate model weights such that the truth is approximated by a composite model ensemble. Again, this is also different from BCMA during which the combined models are a-priori defined by the modeler and not meant to only approximate the truth but to allow for its identification. The latter option to select a model of justified complexity remains as the way to approximate the truth when only one best model is desired. This is achieved by CV and actual information-theoretic ICs. Thereby, if no model can clearly outperform the others during the selection, preliminary CV-based weights can be assigned to the models to handle the uncertainty in selecting one. However, note that this is no actual model combination and therefore should also not be applied as such.

5.5 Summary, Discussion and Conclusion

Model combination comes with many different connotations. Yet, in the context of Bayesian multi-model frameworks, we can only pursue two different goals when combining multiple fully developed models:

- Combining predictive distributions for high predictive coverage in case no process-identification is possible (Bayesian Stacking).
- Combining the models on the process-level, i.e., their forecasts, for process-identification in case the true model is assumed to be a superposition of alternative models (BCMA/BCMS).

Both methods are suitable if the models in the set are assumed to "encircle" the true model. Yet, they follow a prediction-driven (*imitating*) or identification-driven (*following*) spirit, respectively. Together with the model selection (and preceding averaging) frameworks BMS/BMA and Pseudo-BMS/BMA, these two frameworks for rating model combinations resemble four distinct options of Bayesian multi-modelling. In each case, averaging of models has a different meaning.

Guidance regarding which framework to choose starts with the single question of whether there is an "Identifiable (quasi-)true model?". If the answer is yes, we

can try to identify it. If the answer is no, predictive approximation is the only meaningful option. Looking at the set of available models, it is then up to the modeller to decide whether a selection framework or combination framework (model space enrichment) shall be applied in order to achieve either process-identification or predictive approximation of the truth. It may be beneficial to support this decision by a preceding analysis of the model setting regarding the interrelation of the models in the set and toward the true model (see Annan and Hargreaves, 2010; Sanderson et al., 2015). This allows to estimate whether it can be assumed that the competitors "encircle", individually approximate or one might even match the true model.

Before applying a multi-model framework, further pre-processing is advisable: Modellers should account for interdependencies of models in a multi-model ensemble, because conceptually coherent models might not result in better coverage of model space but might only add redundancy within the model combination (see, e.g., George et al., 2010; Sanderson et al., 2015). A potential solution to this issue is given by so-called dilution priors as suggested for BMS/BMA (George et al., 2010, and references therein): Redundancy is quantified by a correlation matrix between models (e.g., Garthwaite and Mubwandarikwa, 2010) or a distance metric between predictive distributions (George et al., 2010). Then, accordingly, prior model weights are adjusted, e.g., away from uniform weights before updating. As result, a model that is strongly redundant in comparison to the other members in the set receives a lower model weight. Similarly, this is imaginable for Pseudo-BMS/BMA.

Contrarily to selecting a single best model or combining several competitors in a multi-model framework, there are other modelling paradigms, e.g., as proposed by Neal (1996): "Sometimes a simple model will outperform a more complex model... Nevertheless, I believe that deliberately limiting the complexity of the model is not fruitful when the problem is evidently complex. Instead, if a simple model is found that outperforms some particular complex model, the appropriate response is to define a different complex model that captures whatever aspect of the problem led to the simple model performing well." Gelman et al. (2014) follow this spirit and suggest to construct an expanding and fully-encompassing model that "spans" an entire area of model space $\mathcal{M}$ and contains individual models (that would build a set $\boldsymbol{M}$) as special cases. Yet, it remains questionable whether this approach is suitable for process-identification and whether it can be directly transferred to disciplines like hydrosystem modelling where various different types and fidelities of models are employed (Höge et al., 2019). Nonetheless, in particular if features of the underlying system are (yet) unknown, it resembles a pragmatic approach to increase predictive power.

6 Conclusion & Outlook

Looking back at the introduction of this thesis, conceptual uncertainty between multiple alternative models poses a fundamental problem in every modelling task. I highlighted three research questions (RQ) to systematically address this challenge based on their answers:

Any model rating score requires a proper implementation of Occam's razor: For explanatory or predictive model selection, it is insufficient to only base the rating on a measure of model fit to data like from a plain error metric. Answering RQ 1, I elucidated how the law of parsimony is enforced by a certain data-dependent representation of model complexity within a model rating score, i.e., as Occam's razor. I analysed various model selection criteria that resemble decompositions of such scores. Thereby, I elicited that the model complexity representation within model rating scores defines whether and under which premisses a true model can be identified (consistent model selection) or only approximated. A best model can therefore only be interpreted as best in the respective realm. It is not necessary to generally force our notion of complexity in a strictly defined meaning. But it is necessary to know that when using a certain model rating score, model complexity is rated in a certain way. For model selection, I deduced that asking two principal questions leads to the appropriate class of model selection that employs the suitable model complexity interpretation. The first question distinguishes whether or not a true model can be found. The second question differentiates if uncertainty shall be addressed based on Bayesian probability theory. Within each class, it remains modelling-task specific, whether an implicit evaluation of the model rating score like a marginalized likelihood or its explicit approximation via information criteria is more applicable. This depends on, e.g., computational demand of the models and on underlying assumptions of the criteria about the distributions of data and model predictions. A potential next step would be to couple my classification scheme to a diagnostic system that checks whether the assumptions made by a certain criterion are fulfilled by the modelling task at hand, e.g., based on the numerical samples from model calibration. This system can help to reduce computational demand whenever assumptions are fulfilled. Then, the criterion can be used as "short-cut" to obtain the respective model score instead of full likelihood marginalization or cross-validation. At the same time, it prevents the unjustified application of criteria and misleading results. I am convinced that a lot of misinterpretation of model rating results can be avoided, when systems are used that assist modellers in correctly applying these model rating tools.

A multi-model framework can only be successfully applied relative to the model setting of the modelling task: Every multi-model framework assigns a weight to

each model that expresses conceptual uncertainty. Yet, I advise to keep in mind that conceptual uncertainty has different meanings depending on the underlying $\mathcal{M}$-setting. Answering RQ 2, I demonstrated that the application of multi-model frameworks in the wrong setting leads to misleading conclusions. Only in the $\mathcal{M}$-closed setting, BMS/BMA can reliably be applied and a model weight refers to the probability of being the true model. Outside of $\mathcal{M}$-closed, Pseudo-BMS/BMA and Bayesian Stacking allow for approximating the truth with predictive pdfs of per se wrong models. The former framework therefore searches one best predictive model and expresses preference by model weights, while the latter searches a weighted average of predictive distributions and expresses shares by model weights. For both predictive frameworks, I illustrated the importance of accounting for inferential uncertainty of model weights because seen data might be only an insufficient proxy for unseen data. Hence, I recommend to apply techniques like the Bayesian Bootstrap to counteract erroneous model weights. This technique comes with small additional computational costs but yields trustworthy model weights under limited sample size. As future step, I suggest to also apply this technique to numerical parameter samples when evaluating a marginalized likelihood like BME. In case the numeric evaluation of a certain model is computationally very demanding, Bayesian Bootstrapping can account for insufficient sampling of the parameter space under a limited computational time budget, yielding, e.g., Bayesian Model Selection under time constraints.

The choice of the appropriate multi-model framework emerges naturally when starting from the philosophical perspective on the modelling task: There is no unified take on conceptual uncertainty and no single framework that is always applicable. Hence, I suggest to interpret conceptual uncertainty strictly from the perspective on either branch of the proposed guide to allow for its adequate handling or reduction. Answering RQ 3, I have elaborated which (Bayesian) multi-model frameworks can be distinguished and under which circumstances each of them is properly employed. I disentangled different notions of model combination which showed why only Bayesian Stacking and the introduced BCMS/BCMA are multi-model frameworks in terms of combining fully developed models. Averaging of models for model combination on the level of predictive distributions or forecasts can only be accomplished by these two Bayesian approaches, respectively. Juxtaposed are the two philosophically distinct model selection frameworks, BMS/BMA and Pseudo-BMS/BMA, for which model averaging is a preliminary compromise. Rather than deciding whether one wants to select or combine models in a set in advance, I recommend to follow the proposed guide that decides for the modeller which approach is most promising. The development and discussion of the guide showed that the usage of multiple models opens numerous options to increase ex-

planatory or predictive power, also over conceptual boundaries. A potential future step would be, e.g., to extend the proposed guide by so-called recursive Bayesian estimators (e.g., particle filter, Kalman filter, etc.) as alternatives or complements to model selection and combination. For instance, in the $\mathcal{M}$-open setting where a true model cannot even be conceptualized, sequential data-assimilation via such recursive Bayesian estimators is a beneficial enhancement to increase predictive reliability.

Above all, this thesis emphasizes the Bayesian paradigm - and this is the core conclusion - to explicitly state every aspect of statistical modelling: The choice of a certain model type to address a modelling task at hand, the information about its parameters and also the model setting in which the respective modelling goal shall be achieved - all of these are based on assumptions that require justification. Following the Bayesian paradigm, I suggest to explicitly name and discuss all assumptions made when applying Bayesian multi-model frameworks. Based on this, full inference can be conducted and reliable conclusions can be drawn.

The gained insights and drawn conclusions from my thesis allow us to explore new avenues in future research:

- The application of Bayesian multi-model frameworks to large-scale systems: How can we assure, that the frameworks keep what they promise when used on the field scale? There, we often have complex and computationally expensive models supported by only scarce data of and limited insight to the underlying system. Bayesian methods therefore require a special focus on assigning (intersubjective or even objective) priors and probabilistic treatment of model ranking results, e.g., via the Bayesian Bootstrap.

- A systematic approach to (Bayesian) model reduction: How can we use methods and insights from model ranking and selection to obtain simpler but effective models? Under limited data, an (underdetermined) model can be reduced to the number of functional terms that are actually identified or supported by the data, i.e. a "lower-dimensional effective theory" (Mattingly et al., 2018).

- Model rating scores as objectives in optimal design of experiments (Nowak and Guthke, 2016): How can we anticipate the optimal data required to inform the marginalized likelihoods for consistent (BMS/BMA) or non-consistent (Pseudo-BMS/BMA) model selection? Just like model selection cannot be properly performed without consideration of the model setting, optimal design can be developed further to pursue process-identification in the $\mathcal{M}$-closed setting or high predictive power outside of it.

Further points were discussed by Höge et al. (2019):

- A systematic coverage of model space: How can we assure that the considered set of models is an adequate representation of model space? Model (combination) probabilities are always conditional on that set which necessitates approaches to measure the currently covered model space and to assign consistent prior model probabilities.

- A solid treatment of measurement and model structural errors in multi-model approaches: How can we adequately describe our assumptions about model-specific structural errors and model-independent measurement errors? The appropriateness of commonly used likelihood functions is questionable and more realistic descriptions, e.g. in the form of statistical error models, are needed.

- A guideline for model development: Can the knowledge gained in the evaluation of (yet non-true) models be utilized for creating enhanced model candidates? A structured scheme of how to isolate the strong parts of a model and transfer it to another model with "complementary strong" features would boost model development and system understanding.

Höge et al. (2019) state that "we hope to further strengthen the utility of Bayesian methods in the face of conceptual uncertainty ... by directing their use into the right channels." This thesis is my contribution.

Appendix

A Numerical Methods for Bayesian Inference

Generally, numerical samples of a distribution can be obtained by systematic methods like Latin-Hypercube-Sampling, random methods like Monte Carlo, or hybrid schemes like so-called orthogonal sampling (McKay et al., 1979).

A.1 Importance Sampling

Let us take $p^*(\Theta)$ (with Θ as one-dimensional variable for simplicity) as our distribution of interest we want to sample from, e.g., a prior parameter distribution. If $p^*(\Theta)$ can be directly sampled, it is trivial to obtain representative numerical samples, e.g., by plain Monte Carlo sampling (cf. Section 2.4.1), i.e., independent random (direct) drawings from $p^*(\Theta)$. Sometimes, however, it might be only possible to evaluate $p^*(\Theta_i)$ for a certain Θ_i but the entire $p^*(\Theta)$ is not fully tractable. If this is the case and another pdf $g(\Theta)$ in the same space can be sampled more easily, so-called importance sampling (IS; Hammersley and Handscomb, 1964) allows to obtain statistical samples from $p^*(\Theta)$ (indirectly), as visualized in Figure 23.

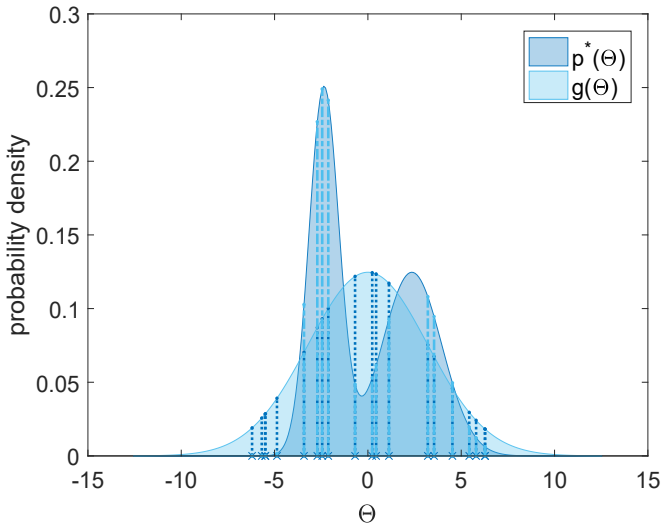


Figure 23: Importance sampling illustration: The (complex) distribution of interest $p^*(\Theta)$ cannot be directly sampled. The (simple) importance distribution $g(\Theta)$ is used to obtain samples from Θ, indicated by vertical dotted lines. The sample weight w_i for each Θ_i is calculated as ratio of the two densities at Θ_i, indicated by the respective dashed and dotted lines.

The idea behind importance sampling is to actually sample from the known (ideally simple) $g(\Theta)$, i.e., the so-called importance distribution (Gelman and Meng, 1998). Then, the samples are weighted by the ratio of densities each sampled Θ_i obtains in $g(\Theta)$ and $p^*(\Theta)$. The weighted samples now represent the target pdf $p^*(\Theta)$. The weight of each sampled Θ_i is $w_i = \frac{p^*(\Theta_i)}{g(\Theta_i)}$. Statistically, this weighting is a correction for sampling from $g(\Theta)$ instead of $p^*(\Theta)$, and direct sampling resembles the special case $g(\Theta) = p^*(\Theta)$ with $w_i = 1$. If $g(\Theta)$ is a proper normalized pdf, the normalizing constant $\mathcal{Z}$ of $p^*(\Theta)$ can be expressed and approximated by:

$$\mathcal{Z} = \int p^*(\Theta)\mathrm{d}\Theta = \int \frac{p^*(\Theta)}{g(\Theta)} g(\Theta)\mathrm{d}\Theta = \mathrm{E}_g\left[\frac{p^*(\Theta)}{g(\Theta)}\right] \approx \sum_{i=1}^{N} \frac{p^*(\Theta_i)}{g(\Theta_i)} \tag{45}$$

In the approximation, N is the number of numerical samples. However, often, it is not necessary to evaluate normalizing constants on an absolute scale but to evaluate relative ratios between two (not necessarily normalized) target pdfs $p_1^*(\Theta)$ and $p_0^*(\Theta)$, e.g. see Bayes Factors in Section 2.3.1. Then, with importance distribution $g(\Theta)$, the corresponding ratio writes as (see, e.g., Lartillot and Philippe, 2006):

$$\frac{\mathcal{Z}_1}{\mathcal{Z}_0} = \frac{\mathrm{E}_g^1\left[\frac{p_1^*(\Theta)}{g(\Theta)}\right]}{\mathrm{E}_g^0\left[\frac{p_0^*(\Theta)}{g(\Theta)}\right]} \approx \frac{\sum_{i=1}^{N} \frac{p_1^*(\Theta_i)}{g(\Theta_i)}}{\sum_{i=1}^{N} \frac{p_0^*(\Theta_i)}{g(\Theta_i)}} \tag{46}$$

N is the number of numeric samples. Equation 45 is therefore only a special case of equation 46 for $p_0^*(\Theta) = g(\Theta)$.

Equation 46 nicely compresses common sampling-based estimators for the marginalized likelihoods in BMS/BMA and Pseudo-BMS/BMA: The assignment of target distributions $p_1^*(\Theta)$ and $p_0^*(\Theta)$ defines whether BMS/BMA or Pseudo-BMS/BMA is conducted. The assignment of $g(\Theta)$ results in a certain Phythagorean mean (arithmetic or harmonic) as estimator for BME or $elpd_{LOO}$, respectively.

- In BMS/BMA the densities of interest are defined as $p_1^*(\Theta) = p(\boldsymbol{D}|\Theta)p(\Theta)$ and $p_0^*(\Theta) = p(\Theta)$. With the prior $p(\Theta)$ being contained in both, weights can be defined as $w_i = \frac{p(\Theta_i)}{g(\Theta_i)}$ and the BME is generally given by:

$$p(D) = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} = \frac{\mathrm{E}_g^1\left[\frac{p(\boldsymbol{D}|\Theta)p(\Theta)}{g(\Theta)}\right]}{\mathrm{E}_g^0\left[\frac{p(\Theta)}{g(\Theta)}\right]} \approx \frac{\sum_{i=1}^{N} \frac{p(\boldsymbol{D}|\Theta_i)p(\Theta_i)}{g(\Theta_i)}}{\sum_{i=1}^{N} \frac{p(\Theta_i)}{g(\Theta_i)}} = \frac{\sum_{i=1}^{N} p(\boldsymbol{D}|\Theta_i)w_i}{\sum_{i=1}^{N} w_i}$$

– Using samples from $g(\Theta) = p(\Theta)$ yields $w_i = \frac{p(\Theta_i)}{p(\Theta_i)} = 1$ and results in the arithmetic mean estimator (AME; as used in Chapter 4):

$$p(D) \approx \frac{1}{N} \sum_{i=1}^{N} p(D|\Theta_i)$$

– Using samples from $g(\Theta) = p(\mathrm{D}|\Theta)p(\Theta)$ yields $w_i = [p(D|\Theta_i)]^{-1}$ and results in the harmonic mean estimator (HME; Newton and Raftery, 1994):

$$p(D) \approx \left[\frac{1}{N} \sum_{i=1}^{N} [p(D|\Theta_i)]^{-1} \right]^{-1}$$

The AME is computationally expensive but, due to the law of large numbers, the most reliable estimator for $p(D)$ (Schöniger et al., 2014). The HME is computationally less expensive because it can be evaluated based on posterior samples from advanced sampling methods like MCMC. However, it could be shown to be a very unreliable estimator for $p(D)$ because a harmonic mean is always very sensitive to the variance of the averaged quantity (see, e.g., Neal, 1996; Vehtari et al., 2017). Note, that, outside of the $\mathcal{M}$-closed setting, the posterior might be a bad importance distribution for the prior that actually needs to me marginalized out for obtaining BME.

- Alternatively, in Pseudo-BMS/A , the target pdfs are assigned as $p_1^*(\Theta) = p(D_o|\Theta)p(\Theta|\boldsymbol{D}_\varnothing)$ and $p_0^*(\Theta) = p(D_o|\boldsymbol{D}_\varnothing)$ and the weights are $w_i = \frac{p(\Theta_i|\boldsymbol{D}_\varnothing)}{g(\Theta_i)}$, which yields:

$$p(D_o|\boldsymbol{D}_\varnothing) = \frac{\mathcal{Z}_1}{\mathcal{Z}_0} = \frac{\mathrm{E}_g^1 \left[\frac{p(D_o|\Theta)p(\Theta|\boldsymbol{D}_\varnothing)}{g(\Theta)} \right]}{\mathrm{E}_g^0 \left[\frac{p(\Theta|\boldsymbol{D}_\varnothing)}{g(\Theta)} \right]} \approx \frac{\sum_{i=1}^{N} p(D_o|\Theta_i) w_i}{\sum_{i=1}^{N} w_i}$$

– Using samples from $g(\Theta) = p(\Theta|\boldsymbol{D}_\varnothing)$ yields $w_i = \frac{p(\Theta_i|\boldsymbol{D}_\varnothing)}{p(\Theta_i|\boldsymbol{D}_\varnothing)} = 1$ and results in the arithmetic mean estimator (AME):

$$p(D_o|\boldsymbol{D}_\varnothing) \approx \frac{1}{N} \sum_{i=1}^{N} p(D_o|\Theta_i)$$

– Using samples from $g(\Theta) = p(\Theta|\boldsymbol{D})$ yields $w_i = [p(D_o|\Theta_i)]^{-1}$ and results in the harmonic mean estimator (HME):

$$p(D_o|\boldsymbol{D}_\varnothing) \approx \left[\frac{1}{N} \sum_{i=1}^{N} [p(D_o|\Theta_i)]^{-1} \right]^{-1}$$

Usually, when estimating $elpd_{LOO}$ as $\sum_{o=1}^{N_s} p(D_o|\boldsymbol{D}_{\varnothing})$, not each LOO-posterior is sampled individually. Normally, the full posterior is sampled and reweighted. The underlying assumption is that the full posterior is very close to each LOO-posterior and therefore serves as suitable importance distribution. Nonetheless, the HME is an instable estimator that suffers from high variance of the evaluated $p(D_o|\Theta_i)$. Hence, stabilization schemes exist, e.g., the so-called pareto-smoothed importance sampling (PSIS; Vehtari et al., 2017) that automatically provides additional key values to rate the reliability of the stabilization.

A contrasting summary between BMS/BMA and Pseudo-BMS/BMA regarding model weights, marginalized likelihood and all respective estimators (importance sampling-based and IC-based, see Section 2.5) is given in Table 8.

Table 8: Contrasting BMS/BMA and Pseudo-BMS/BMA: Summary of model weights, the respective marginalized likelihoods BME and $elpd_{LOO}$ and numerical estimators for multi-dimensional data $\boldsymbol{D}$ and parameters $\boldsymbol{\Theta}$: importance sampling (IS)-based arithmetic (AME) and harmonic (HME) mean estimators for importance distribution $g(\boldsymbol{\Theta})$; information criteria (IC) as approximative estimators. Estimators are indicated by $\widehat{\cdot}$, the number of observations is N_s and the number of numeric samples is N.

	BMA/BMS	Pseudo-BMA/BMS
model weight	$w_m = \frac{p(\boldsymbol{D}\|M_m)p(M_m)}{\sum_{k=1}^{N_M} p(\boldsymbol{D}\|M_k)p(M_k)}$	$w_m = \frac{p(\boldsymbol{D'}\|\boldsymbol{D},M_m)}{\sum_{k=1}^{N_M} p(\boldsymbol{D'}\|\boldsymbol{D},M_k)}$
core quantity	$p(\boldsymbol{D}) = \int p(\boldsymbol{D}\|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}$ (prior predictive density - evaluated at $\boldsymbol{D}$)	$p(\boldsymbol{D'}\|\boldsymbol{D}) = \int p(\boldsymbol{D'}\|\boldsymbol{\Theta})p(\boldsymbol{\Theta}\|\boldsymbol{D})\mathrm{d}\boldsymbol{\Theta}$ (posterior predictive density - estimated for $\boldsymbol{D'}$)
- proxy		$elpd_{LOO} = \sum_{o=1}^{N_s} \ln p(D_o\|\boldsymbol{D}_{\varnothing})$ with $p(D_o\|\boldsymbol{D}_{\varnothing}) = \int p(D_o\|\boldsymbol{\Theta})p(\boldsymbol{\Theta}\|\boldsymbol{D}_{\varnothing})\mathrm{d}\boldsymbol{\Theta}$
IS estimators	$\widehat{p(\boldsymbol{D})} = \frac{\sum_{i=1}^{N} p(\boldsymbol{D}\|\boldsymbol{\Theta}_i)w_i}{\sum_{i=1}^{N} w_i}$ with $w_i = \frac{p(\boldsymbol{\Theta}_i)}{g(\boldsymbol{\Theta}_i)}$	$\widehat{elpd}_{LOO} = \sum_{o=1}^{N_s} \ln \frac{\sum_{i=1}^{N} p(D_o\|\boldsymbol{\Theta}_i)w_i}{\sum_{i=1}^{N} w_i}$ with $w_i = \frac{p(\boldsymbol{\Theta}_i\|\boldsymbol{D}_{\varnothing})}{g(\boldsymbol{\Theta}_i)}$
- AME	$\widehat{p(\boldsymbol{D})} = \sum_{i=1}^{N} p(\boldsymbol{D}\|\boldsymbol{\Theta}_i)$ for $g(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta})$	$\widehat{elpd}_{LOO} = \sum_{o=1}^{N_s} \ln \sum_{i=1}^{N} p(D_o\|\boldsymbol{\Theta}_i)$ for $g(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}\|\boldsymbol{D}_{\varnothing})$
- HME	$\widehat{p(\boldsymbol{D})} = \left[\sum_{i=1}^{N} p(\boldsymbol{D}\|\boldsymbol{\Theta}_i)^{-1}\right]^{-1}$ for $g(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}\|\boldsymbol{D})$	$\widehat{elpd}_{LOO} = \sum_{o=1}^{N_s} \ln \left[\sum_{i=1}^{N} p(D_o\|\boldsymbol{\Theta}_i)^{-1}\right]^{-1}$ for $g(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}\|\boldsymbol{D})$
IC	$\ln \widehat{p(\boldsymbol{D})}_{WBIC} \approx -\frac{1}{2}\mathrm{WBIC}$ $\ln \widehat{p(\boldsymbol{D})}_{KIC} \approx -\frac{1}{2}\mathrm{KIC}$ $\ln \widehat{p(\boldsymbol{D})}_{BIC} \approx -\frac{1}{2}\mathrm{BIC}$	$\widehat{elpd}_{WAIC} \approx -\frac{1}{2}\mathrm{WAIC}$ $\widehat{elpd}_{DIC} \approx -\frac{1}{2}\mathrm{DIC}$ $\widehat{elpd}_{AIC} \approx -\frac{1}{2}\mathrm{AIC}$

A.2 Power-Posterior and Thermodynamic Integration

Besides the AME and HME, also a geometric mean approach for estimating a marginalized likelihood exists. Therefore, the concept of so-called power-posterior distributions (Friel and Pettitt, 2008) has to be considered, focusing on the estimation of BME.

A.2.1 Power-Posterior Distributions

A power-posterior is defined over a so-called geometric path with the exponent $\beta \in [0, 1]$ (Gelman and Meng, 1998) between prior and (not normalized) posterior as:

$$q_\beta(\boldsymbol{\Theta}) = \underbrace{p(\boldsymbol{\Theta})}_{prior}^{(1-\beta)} \underbrace{(p(\boldsymbol{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}))^\beta}_{not\ norm.\ post.} = p(\boldsymbol{D}|\boldsymbol{\Theta})^\beta p(\boldsymbol{\Theta}) \tag{47}$$

The normalizing constant of the power-posterior for a certain β is:

$$\mathcal{Z}_\beta = \int q_\beta(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \tag{48}$$

By tempering the likelihood, $q_\beta(\boldsymbol{\Theta})$ transitions between the prior ($\beta = 0$) and the unnormalized posterior ($\beta = 1$) distribution. The respective normalizing constants are $\mathcal{Z}_0 = 1$ and $\mathcal{Z}_1 = \text{BME}$. For an information-theoretic interpretation of the tempered distributions refer to Vitoratou and Ntzoufras (2017).

A.2.2 Thermodynamic Integration

Thermodynamic integration in the context of Bayesian inference refers to obtaining the logarithmic BME $\lambda = \ln p(\boldsymbol{D})$ based on power-posteriors and Equation 48. The terminology clearly indicates the original motivation which was to estimate the free energy of a physical system which is the energy density in the system integrated over all its states, i.e. a marginalized quantity (Gelman and Meng, 1998). Hence, BME is sometimes also referred to as Bayesian free-energy (Watanabe, 2010). The thermodynamic interpretation of the exponent $\beta \in [0, 1]$ is therefore an inverse temperature T:

$$\underbrace{1 = \frac{1}{T_{min}}}_{T_{min}=1} \geq \beta \geq \frac{1}{T_{max}} \quad \text{with} \quad \lim_{T_{max}\to\infty} \frac{1}{T_{max}} = 0$$

Generally expressed, ratios of normalizing constants $\mathcal{Z}$ of two (physical or statistical) densities can be computed with thermodynamic integration by constructing a path in between them. Thermodynamic integration is another name for path sampling (Gelman and Meng, 1998) that recognizes its motivation from statistical physics. The logarithmic BME writes as (e.g., Lartillot and Philippe, 2006):

$$\lambda = \int_0^1 \frac{\partial \ln \mathcal{Z}_\beta}{\partial \beta} \mathrm{d}\beta$$

with

$$\begin{aligned}\frac{\partial \ln \mathcal{Z}_\beta}{\partial \beta} &= \frac{1}{\mathcal{Z}_\beta} \frac{\partial \mathcal{Z}_\beta}{\partial \beta} = \frac{1}{\mathcal{Z}_\beta} \frac{\partial}{\partial \beta} \int q_\beta(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} = \frac{1}{\mathcal{Z}_\beta} \int \frac{\partial q_\beta(\boldsymbol{\Theta})}{\partial \beta} \mathrm{d}\boldsymbol{\Theta} \\ &= \int \frac{1}{q_\beta(\boldsymbol{\Theta})} \frac{\partial q_\beta(\boldsymbol{\Theta})}{\partial \beta} \frac{q_\beta(\boldsymbol{\Theta})}{\mathcal{Z}_\beta} \mathrm{d}\boldsymbol{\Theta} = \int \frac{\partial \ln q_\beta(\boldsymbol{\Theta})}{\partial \beta} p_\beta(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta}\end{aligned}$$

yields

$$\lambda = \int_0^1 \int \frac{\partial \ln q_\beta(\boldsymbol{\Theta})}{\partial \beta} p_\beta(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \, \mathrm{d}\beta = \int_0^1 \mathrm{E}_{p_\beta} \left[\frac{\partial \ln q_\beta(\boldsymbol{\Theta})}{\partial \beta} \right] \mathrm{d}\beta \tag{49}$$

For the geometric path between $q_0(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta})$ and $q_1(\boldsymbol{\Theta}) = p(\boldsymbol{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})$ (see Equation 47) this is:

$$\begin{aligned}\lambda &= \int_0^1 \mathrm{E}_{p_\beta} \left[\frac{\partial \ln \left(q_0(\boldsymbol{\Theta})^{(1-\beta)} \, q_1(\boldsymbol{\Theta})^\beta \right)}{\partial \beta} \right] \mathrm{d}\beta = \int_0^1 \mathrm{E}_{p_\beta} \left[\ln \frac{q_1(\boldsymbol{\Theta})}{q_0(\boldsymbol{\Theta})} \right] \mathrm{d}\beta \\ &= \int_0^1 \mathrm{E}_{p_\beta} \left[\ln p(\boldsymbol{D}|\boldsymbol{\Theta}) \right] \mathrm{d}\beta\end{aligned} \tag{50}$$

In practice, the thermodynamic integral is usually solved over a latter of discrete temperatures β (Friel and Pettitt, 2008; Friel et al., 2013). By dampening the likelihood function, thermodynamic integration is a suitable method in particular if the likelihood function causes the posterior to be a "challenging" distribution, e.g., being multi-modal. Further, thermodynamic integration is a suitable method if $q_0(\boldsymbol{\Theta})$ and $q_1(\boldsymbol{\Theta})$ barely overlap (Liu et al., 2016). However, if this is the case, again, it is questionable whether consistent model rating via BME is appropriate (cf. Section 2.2).

A.2.3 Related "Tempered" Methods

Another approach for estimating the BME, that is also based on power posteriors, is called stepping-stone method (Li et al., 2010). Rather than a continuous integration over β, this method estimates the BME ($\mathcal{Z}_1$) in a step-wise manner for a sequence of discrete values for $\beta \in [0, 1]$: Between two sequential values of β, the ratio of respective normalizing constants ("tempered" Bayes factors) is evaluated, starting from $\mathcal{Z}_0$. Then, the telescope product of these ratios yields $\mathcal{Z}_1/\mathcal{Z}_0$, i.e. the BME. In statistics, this step-wise evaluation of normalizing constants is known as bridge sampling (Gelman and Meng, 1998) and resembles importance sampling over several stages.

A "thermodynamic" information criterion is the so-called Watanabe-Bayesian IC (WBIC). The WBIC estimates BME based on samples from a power posterior for a distinct (optimal) β^* (Watanabe, 2013). WBIC is the consistent counterpart to WAIC (Section 2.5.3) and both require theoretically only one MCMC chain to be evaluated for sampling the (tempered) posterior (Friel and Wyse, 2012).

A.3 Alternative Methods

A popular sampling-based alternative for Bayesian inference is so-called nested sampling (Skilling, 2004). In nested sampling, the marginalization over the prior parameter distribution is performed as one-dimensional integral over so-called "prior mass". Using predefined likelihood thresholds, the method integrates over the prior mass for each likelihood interval and sums the results to obtain the marginal likelihood (see, e.g., Schöniger et al., 2014).

In contrast to sampling-based approaches, Variational Inference (e.g., Brodersen et al., 2013; Blei et al., 2017) turns the marginalization problem into an optimization problem: Briefly, a predefined distribution (e.g., a conjugate distribution, see Section 2.4.1) with unknown distributional parameters is fitted as posterior (model) parameter distribution given data $\boldsymbol{D}$. The fitting is done by maximizing a so-called evidence lower bound (which refers to BME) in an information-theoretic sense while minimizing D_{KL} between the corresponding posterior predictive distribution and the true data distribution (Blei et al., 2017). Thereby, Variational Inference works similar to so-called Expectation-Maximization (Dempster et al., 1977) in terms of fitting a predefined distribution to values that are seen as samples from an underlying distribution. Expectation-Maximization yields a maximum likelihood estimate for the best distributional parameters. Variational Inference further allows to include prior (model parameter) knowledge and is therefore suitable for Bayesian inference.

A.4 Numerical Techniques

Generally spoken, numerical techniques in Bayesian inference address Bayesian updating and marginalization by employing optimization (see Variational Inference in Section A.3) or sampling algorithms (see MCMC in Section 2.4.1). As a rule of thumb: Optimization-based techniques are comparably fast but require (strong) assumptions about the distributions. Sampling-based techniques typically require no or only weak assumptions about the distributions but need a certain amount of samples to converge to the full distribution and are therefore computationally expensive.

When a certain modelling task in applied modelling prohibits strong assumptions about posterior distributions, Bayesian inference is usually conducted based on numerical samples. As widely applicable and popular tool, I want to give a quick overview of the variety of MCMC. Regarding the basics about MCMC and convergence, numerous sources of information are available, e.g., Andrieu et al. (2003) or Gelman and Rubin (1992). Sophisticated MCMC methods that were developed to address challenges in sampling are shown in Table 9 right next to the respective issue. All of them provide chains of numerical samples to represent the distribution of interest but the ways how these chains are built differ vastly.

Table 9: Sophisticated MCMC techniques: Issues that exacerbate numerical sampling and the respective MCMC method to solve them.

Challenge	Solution
huge dataset	Stochastic MCMC (Simsekli et al., 2016): MCMC with data subsampling
high-dim. space	preconditioned Crank-Nicolson MCMC (Cotter et al., 2013) or similar methods: dimension-independent sampling
slow convergence	Hamiltonian MCMC (e.g. Hoffman and Gelman (2014)): fictitious momentum (Hamiltonian energy operator) for jump proposal
sampling of two spaces of different dimensionality	Reversible Jump MCMC (Green, 1995): MCMC that jumps within and in-between spaces
data-assimilation of incoming data	Sequential MCMC (Yang and Dunson, 2013): sampling from a sequence of distributions (e.g. dynamic posterior over time)
multi-modality	Parallel Tempering MCMC (Earl and Deem, 2005): chains with different temperatures

Some of these numerical techniques are applicable in direct correspondence to statistical methods. A straight-forward example is the combination of thermodynamic integration with Parallel Tempering MCMC (MCMC-PT; e.g., Vousden

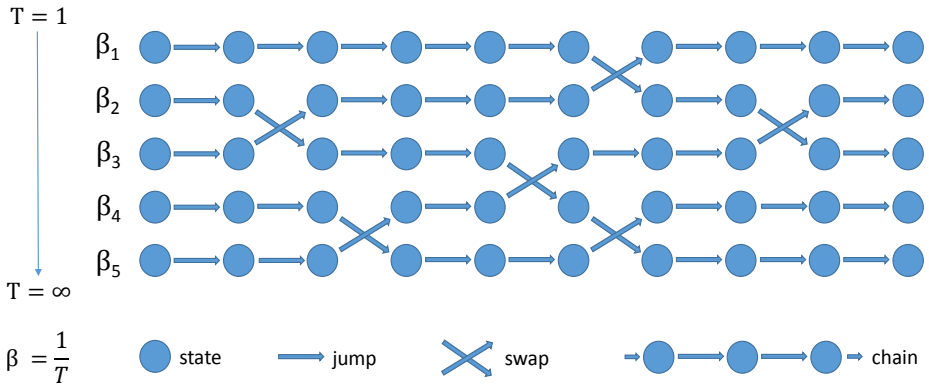


Figure 24: Parallel Tempering MCMC scheme: Example with five parallel chains at different (inverse) temperatures β and the core concept of jumps and swaps between states in the sampling space.

et al., 2016) as named last in Table 9. Hence, I want to give a quick qualitative introduction to MCMC-PT, a.k.a. population MCMC (Calderhead and Girolami, 2009): MCMC-PT uses chains at different inverse temperatures β to "explore and exploit" the distribution of interest, e.g., a posterior parameter distribution:

- Hot chains (small β) *explore* the (e.g, parameter) space with large jumps. They might leave high probability regions before they are sufficiently sampled but they do not get stuck in local maxima.
- Cold chains (large β) *exploit* the (e.g, parameter) space with small jumps. Once found, they thoroughly sample high probability regions but might get stuck in local maxima.

The core strength of MCMC-PT is: The different tempered chains can communicate with one another and occasionally swap positions. Thus, hot chains that explored high probability regions lead cold chains there for them to exploit these regions in detail. This makes MCMC-PT suitable also for multi-modal distributions (Vousden et al., 2016) as they often appear with overparameterized and underdetermined models. The scheme behind PT is illustrated in figure 24.

A.5 Available Software

An overview about software that provides easy access to numerical Bayesian Inference was assembled for Höge et al. (2019): "Readily available software packages for Bayesian modeling and model evaluation exist (PyMC3 (Salvatier et al., 2016), STAN (Carpenter et al., 2017), WinBUGS (Lunn et al., 2000), JAGS (Plummer

et al., 2003), etc.). However, most of them provide Bayesian CV-based rather than BME-based model evaluation and ranking. The ones that allow for BME evaluation are often sampling algorithms that provide BME as a side-product (emcee (Foreman-Mackey et al., 2013), DREAM (Vrugt et al., 2009), MrBayes (Huelsenbeck and Ronquist, 2001), etc.)."

B Applied Model Complexity Control

B.1 Model Complexity within Selection Criteria: Synthetic Example

In extension to the comparison of the four model selection classes in the Section 3.2, the four complexity representations are evaluated in a synthetic example over increasing sample size. For each class, one representative member is presented:

- Class B1 (Section 2.5.1): Occam factor OF as complexity representation in the KIC (Equation 27).
- Class B0 (Section 2.5.2): Geometric complexity GC as complexity representation in the MDL (Equation 29).
- Class A1 (Section 2.5.3): Effective number of parameters N_p^* as complexity representation in the WAIC (Equation 36).
- Class A0 (Section 2.5.4): Model degrees of freedom DoF as complexity representation in the EPE (Equation 39).

For illustration, four simple models are used, two of which are linear (polynomial) and two are non-linear (power-law and sine). The linear models are termed L1/L2 and the non-linear models NL1/NL2, respectively. They are summarized and contrasted to the data-generating process (DGP) in Table 10.

Table 10: Synthetic truth (DGP) and models for evaluating model complexity measures: Off-diagonal entries in parameter covariance matrix are always $\Sigma_{i,j\neq i} = 0$

Model	Equation	Parameter	
DGP	$\Theta_1\mathbf{x} + \Theta_2 + \boldsymbol{\epsilon}$	$\hat{\overline{\boldsymbol{\Theta}}} = [0.95; 0.05]$	$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$ with $\sigma_\epsilon = 0.3$
L1	$\Theta_1\mathbf{x} + \Theta_2$	$\overline{\boldsymbol{\Theta}} = [1; 0]$	$\Sigma_{1,1} = 0.25$, $\Sigma_{2,2} = 0.25$
L2	$\Theta_1\mathbf{x}^3 + \Theta_2\mathbf{x}^2 ...$ $+\Theta_3\mathbf{x} + \Theta_4$	$\overline{\boldsymbol{\Theta}} = [-0.0025; 0.05;$ $0.8; -0.1]$	$\Sigma_{1,1} = 0.0025$, $\Sigma_{2,2} = 0.025$, $\Sigma_{3,3} = 0.25$, $\Sigma_{4,4} = 0.25$
NL1	$\Theta_1\mathbf{x}^{\Theta_2}$	$\overline{\boldsymbol{\Theta}} = [1.1; 0.9]$	$\Sigma_{1,1} = 0.0625$, $\Sigma_{2,2} = 0.01$
NL2	$\Theta_1 \sin(\Theta_2\mathbf{x} + \Theta_3) ...$ $+\Theta_4$	$\overline{\boldsymbol{\Theta}} = [6; 5.75; -1;$ $5.25]$	$\Sigma_{1,1} = 0.0625$, $\Sigma_{2,2} = 0.0625$, $\Sigma_{3,3} = 0.01$, $\Sigma_{4,4} = 0.0625$

B.1.1 Numerical Implementation

With the DGP, a dataset of in total 500 data points was generated that was then subdivided in nested datasets of sizes 5, 10, 30, 50, 100, 300 and 500 points. For

each sample size, posterior parameter samples for the four simple models are generated. For the linear models with two (L1) and four (L2) parameters, 50,000 and 250,000 posterior prediction samples from the analytic solution for the posterior were drawn, respectively. The posterior samples for the weakly non-linear power-law model (NL1) and the stronger non-linear sine function (NL2) were gained via importance sampling (see Section A.1). The importance sampling distributions were the respective priors from which 250,000 samples were generated for the NL1 and 500,000 samples were drawn for the NL2. Sufficiency of sampling was ensured by evaluating the effective sample size ESS $= \left(\sum_i^{N_{MC}} w_i^2\right)^{-1}$. With the generation of data and the sampling being performed over 50 realizations, the ESS was averaged over all realizations. For the non-linear models, the average ESS was in the order of 10^4 for $N_s = 5$, decreasing to the order of 10^2 for $N_s = 500$.

Two kinds of prior parameter distributions were used for each model: uniform priors and Gaussian priors. The Gaussian priors were assumed to be uncorrelated (see Table 10). The uniform priors were constructed using the respective Gaussian prior means $\pm 2\sqrt{\Sigma_{i,i}}$, i.e. two times the respective standard deviation per parameter. The evaluated model complexity terms for each prior over 50 realizations were shown as Box-Whisker-plots in Figures 25 and 26: with median (dot), quartiles (bars), whiskers (lines) and outliers (crosses).

B.1.2 B-type Model Complexity

The consistent model complexity measures OF and GC are shown as Box-Whisker-plots in Figure 25. The complexity term in the BIC, $\mathcal{C}_{\mathrm{BIC}} = N_p \ln(N_s)/2$, is used as reference.

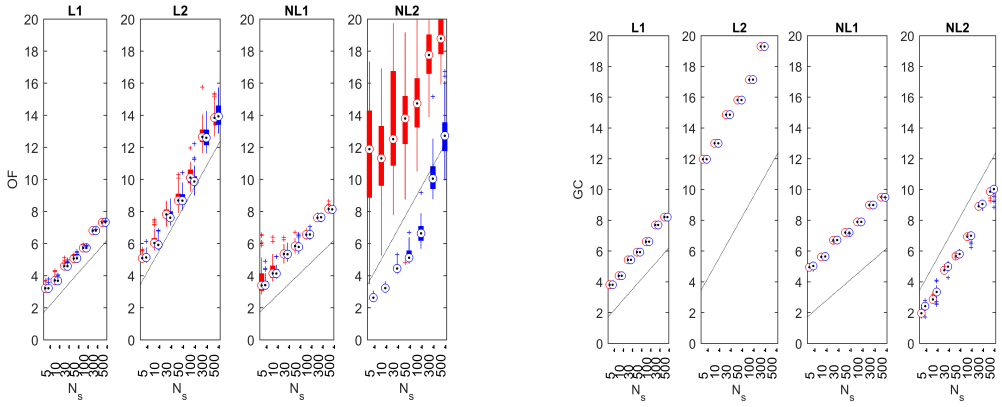
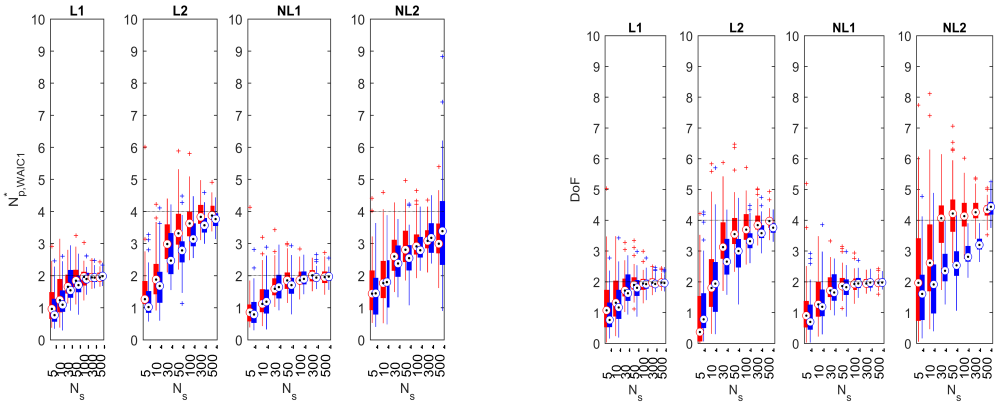


Figure 25: Consistent complexity measures evaluated over growing data size N_s for uniform (red) and Gaussian (blue) parameter prior, respectively: (a) Occam factor (OF) from KIC in class B1; (b) Geometric complexity (GC) from MDL in class B0.

For all models, both complexity representations clearly show the subextensive growth over increasing data size required in consistent model selection, following the trend of $\mathcal{C}_{\mathrm{BIC}}$. For the OF in Figure 25 (a), this resembles the increasing shrinkage of the posterior-prior-ratio. Except for the NL2, there is no significant difference between the assigned priors. For NL2, the Gaussian prior restricts the shrinkage due to the additional prior information about parameters, while in the uniform case, the likelihood function dominates. Due to the strong non-linearity, KIC might be unsuitable for model rating of NL2, which is shown by the large spread of evaluated complexity values. The GC in Figure 25 (b) shows a similar behaviour like OF with increasing sample size N_s. The complexity representations (as distinguishable likelihoods) between the two priors are very similar and, with exception for NL2, always slightly larger than the OF. This might cause differences in model ranking between the two classes.

B.1.3 A-type Model Complexity

The non-consistent model complexity measures N_p^* and DoF are shown as Box-Whisker-plots in Figure 26. The complexity term in the AIC, $\mathcal{C}_{\text{AIC}} = N_p$, is used as reference.



(a) **(b)**

Figure 26: Non-consistent complexity measures evaluated over growing data size N_s for uniform (red) and Gaussian (blue) parameter prior, respectively: (a) Effective number of parameters (N_p^*) from WAIC in class A1; (b) Model degrees of freedom (DoF) from EPE in class A0.

For all models, both complexity representations clearly show the bounded behaviour over increasing data size required in nonconsistent model selection, following the trend of $\mathcal{C}_{\text{AIC}}$. In case of N_p^* in Figure 26(a), this can be interpreted as how many parameters are constrained by the information contained in the data instead of the prior information. This fraction increases over growing data size since the prior becomes negligible and all information comes from the data. In particular, L2 shows that the Gaussian prior contains more information than the truncated uniform one. Again, NL2 is the exception due to the strong nonlinearity: First, the bounding of N_p^* does not seem to coincide with N_p - the number of effective parameters differs from the number of countable parameters. Second, for larger data sizes, the Gaussian prior yields slightly larger N_p^* as median than the uniform prior. This might indicate conflicting information between prior and data, resulting in increased spread of N_p^*. The DoF in Figure 26 (b) shows a similar behaviour like N_p^* over increasing sample size N_s. The complexity for both priors are very similar, with the pattern of additional constraint by the Gaussian prior. Yet, NL2 highlights that the two classes A1 and A0 still differ. All DoF show much larger spread than corresponding N_p^* and in case of a uniform prior, the sensitivity to data perturbations is very large and decreases only over increasing amount of data, with the median converging approximately to N_p. This might

cause differences in model ranking between the two classes.

B.2 Complexity Control in Black-Box Models

When working with black-box models, there is typically no "natural" complexity control by physical conservation laws or expert knowledge about prior parameter distributions, e.g., as in neural networks (see Equation 3). Then, complexity control is enforced by regularization as discussed in Section 3.1, e.g., by assigning prior parameter distributions which yields so-called Bayesian neural networks (e.g. Neal, 1996).

Alternatively, splitting of available data is used to counteract overfitting. Opposed to the traditional splitting into two parts for calibration and validation, the data is split into three parts called: *training* data (for model calibration), *validation* data (for model regularization) and *testing* data (for model ranking) (Friedman et al., 2001). Respective percentages of all available data are, e.g., 50-25-25. This might be confusing because, traditionally, validation data is used for model evaluation and ranking. Now, especially in data-driven machine learning, validation data is used in parallel to the calibration for regularization in order to constrain the parameters – in data-driven modelling, data provides the expert knowledge, too. Calibration is stopped when the performance on the validation data starts to decline which implies increasing risk of overfitting. This procedure is called "early stopping" (e.g., MacKay, 1992). After this within-model complexity control, the remaining test data is used to evaluate the model performance on hold-out data and according ranking between models.

C Analytic Solutions to Marginalized Likelihoods: Gaussian Linear Model

For a linear (polynomial) model with Gaussian parameter distribution and a Gaussian likelihood function, the analytic expression for the logarithmic marginal likelihood (BME) is (e.g., Box and Tiao, 1973):

$$\ln p(\boldsymbol{D})^{linG} = -\frac{1}{2}N_s \ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_{DD}| - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{\Sigma}_{DD}^{-1}\boldsymbol{r} \tag{51}$$

with the column vector of N_s residuals $\boldsymbol{r} = \boldsymbol{D} - \boldsymbol{H}\bar{\boldsymbol{\Theta}}$ between data $\boldsymbol{D}$ and predictions obtained by multiplying the parameter mean $\bar{\boldsymbol{\Theta}}$ matrix of base functions $\boldsymbol{H}$. $\boldsymbol{\Sigma}_{DD}$ can be obtained via linear uncertainty propagation (Schweppe, 1973; Schöniger et al., 2014) with the parameter prior variance-covariance matrix $\boldsymbol{\Sigma}_{\Theta\Theta}$ and the measurement error matrix $\boldsymbol{R}$ via $\boldsymbol{\Sigma}_{DD} = \boldsymbol{H}\boldsymbol{\Sigma}_{\Theta\Theta}\boldsymbol{H}^T + \boldsymbol{R}$.

Formulating Equation 51 for both, $\boldsymbol{D}$ and $\boldsymbol{D}_{\varnothing}$, plugging them into Equation 43, respectively, and rearranging terms yields:

$$elpd_{LOO}^{linG} = -\frac{N_s}{2}\ln(2\pi) + \frac{1}{2}\sum_{o=1}^{N_s}\ln\frac{|\boldsymbol{\Sigma}_{\boldsymbol{D}_{\varnothing}\boldsymbol{D}_{\varnothing}}|}{|\boldsymbol{\Sigma}_{\boldsymbol{DD}}|} - \frac{N_s}{2}\left[\boldsymbol{r}^T\boldsymbol{\Sigma}_{DD}^{-1}\boldsymbol{r} - \frac{1}{N_s}\sum_{o=1}^{N_s}\boldsymbol{r}_{\varnothing}^T\boldsymbol{\Sigma}_{D_{\varnothing}D_{\varnothing}}^{-1}\boldsymbol{r}_{\varnothing}\right] \tag{52}$$

Generally, the more negative $elpd_{LOO}$ is, the smaller the predictive density of future data is expected to be - to the downside of the particular model. A computationally efficient solution to evaluate Equation 52 requires only the inversion of the data variance-covariance matrix to obtain $\boldsymbol{\Sigma}_{DD}^{-1}$:

- The last term can be obtained by exploiting the relation between the inverse of a submatrix (here $\boldsymbol{\Sigma}_{D_{\varnothing}D_{\varnothing}}$) to the inverse of the main matrix (here $\boldsymbol{\Sigma}_{DD}$), proposed by e.g. Juárez-Ruiz et al. (2016): Given a square matrix $\boldsymbol{A}$ of size $n \times n$ and its inverse $\boldsymbol{A}^{-1}$, the inverse of a submatrix $\boldsymbol{A}_{sub}^{-1}$, with $\boldsymbol{A}_{sub}$ obtained by dropping row i and column j from $\boldsymbol{A}$, is

$$\boldsymbol{A}_{sub}^{-1} = \boldsymbol{N} + \frac{\boldsymbol{u}\boldsymbol{v}\boldsymbol{N}}{a_{ij}a_{ji}^{-1}} \tag{53}$$

  $\boldsymbol{N}$ is a submatrix of $\boldsymbol{A}^{-1}$, where row j and column i are dropped; $\boldsymbol{u}$ is column i of $\boldsymbol{A}^{-1}$ without element j; $\boldsymbol{v}$ is row i of $\boldsymbol{A}$ without element j; a_{ij} is the element of $\boldsymbol{A}$ in row i and column j and a_{ji}^{-1} is the element of $\boldsymbol{A}^{-1}$ in row j and column i.

- The ratio of determinants in the second term of equation 52 can likewise be obtained using element a_{ji}^{-1} from $\boldsymbol{A}^{-1}$, since

$$\frac{|\boldsymbol{A}_{sub}|}{|\boldsymbol{A}|} = \frac{1}{(-1)^{i+j}}\, a_{ij}^{-1} \tag{54}$$

Throughout the summations in equation 52, the special case $j = i$ holds and $\boldsymbol{A}_{sub}$ resembles $\boldsymbol{\Sigma}_{D_\varnothing D_\varnothing}$ (with $n = N_s - 1$) for each o.

References

Aho, K., Derryberry, D., and Peterson, T. 2014. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636. doi: 10.1890/13-1452.1.

Ajami, N. K., Duan, Q., and Sorooshian, S. 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, 43(1). doi: 10.1029/2005WR004745.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, pages 267–281.

Akaike, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723. doi: 10.1109/TAC.1974.1100705.

Akaike, H. 1978. A new look at the Bayes procedure. *Biometrika*, 65(1):53–59. doi: 10.1093/biomet/65.1.53.

Albert, C., Künsch, H. R., and Scheidegger, A. 2015. A simulated annealing approach to approximate Bayes computations. *Statistics and computing*, 25(6): 1217–1232.

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. 2003. An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43. doi: 10.1023/A:1020281327116.

Angluin, D. and Smith, C. H. 1983. Inductive inference - theory and methods. *Computing Surveys*, 15(3):237–269. doi: 10.1145/356914.356918.

Annan, J. D. and Hargreaves, J. C. 2010. Reliability of the CMIP3 ensemble. *Geophysical Research Letters*, 37(2).

Arlot, S. and Celisse, A. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Bardsley, W. E., Vetrova, V., and Liu, S. 2015. Toward creating simpler hydrological models: A LASSO subset selection approach. *Environmental Modelling & Software*, 72:33–43. doi: http://dx.doi.org/10.1016/j.envsoft.2015.06.008.

Barron, A., Rissanen, J., and Yu, B. 1998. The minimum description length principle in coding and modeling. *Ieee Transactions on Information Theory*, 44 (6):2743–2760.

Barron, A. R. and Cover, T. M. 1991. Minimum complexity density estimation. *Ieee Transactions on Information Theory*, 37(4):1034–1054.

Bartlett, M. S. 1957. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(3/4):533–534. doi: 10.2307/2332888.

Bates, J. M. and Granger, C. W. 1969. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Bergström, S. and Singh, V. 1995. The HBV model. *Computer models of watershed hydrology.*, pages 443–476.

Bernardo, J. and Smith, A. 1994. *Bayesian Theory. Bayesian Theory.* Wiley, New York. doi: 10.1002/9780470316870.

Betancourt, M. 2015. A unified treatment of predictive model comparison. *arXiv preprint arXiv:1506.02273*.

Beven, K. and Binley, A. 1992. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298.

Bialek, W., Nemenman, I., and Tishby, N. 2001. Complexity through nonextensivity. *Physica A: Statistical Mechanics and its Applications*, 302(1):89–99.

Bishop, C. M. 1995. *Neural networks for pattern recognition. Neural networks for pattern recognition.* Oxford university press, ISBN: 9780198538642.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877.

Blöschl, G. and Sivapalan, M. 1995. Scale issues in hydrological modelling: a review. *Hydrological processes*, 9(3-4):251–290.

Boero, G., Smith, J., and Wallis, K. F. 2011. Scoring rules and survey density forecasts. *International Journal of Forecasting*, 27(2):379–393.

Boisbunon, A., Canu, S., Fourdrinier, D., Strawderman, W., and Wells, M. T. 2014. Akaike's information criterion, Cp and estimators of loss for elliptically symmetric distributions. *International Statistical Review*, 82(3):422–439. doi: 10.1111/insr.12052.

Box, G. and Tiao, G. 1973. Bayesian inference in statistical inference. *Adison-Wesley, Reading, Massachusetts*.

Box, G. E. and Cox, D. R. 1964. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252.

Box, G. E. P. 1976. Science and statistics. *Journal of the american statistical association*, 71(356):791–799. doi: 10.1080/01621459.1976.10480949.

Bozdogan, H. 1990. On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics - Theory and Methods*, 19(1):221–278. doi: 10.1080/03610929008830199.

Bozdogan, H. 2000. Akaike's information criterion and recent developments in information complexity. *Journal of mathematical psychology*, 44(1):62–91. doi: 10.1006/jmps.1999.1277.

Breiman, L. 1996. Stacked regressions. *Machine learning*, 24(1):49–64.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). pages 199–231. doi: 10.1214/ss/1009213726.

Britz, D. 2015. Implementing a neural network from scratch in Python – an introduction. *URL http://www.wildml.com/2015/09/implementing-a-neural-network-from-scratch/*.

Brodersen, K. H., Daunizeau, J., Mathys, C., Chumbley, J. R., Buhmann, J. M., and Stephan, K. E. 2013. Variational Bayesian mixed-effects inference for classification studies. *NeuroImage*, 76:345–361. doi: http://dx.doi.org/10.1016/j.neuroimage.2013.03.008.

Burnham, K. P. and Anderson, D. R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. volume 2. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, ISBN: 0387953647.

Burnham, K. P. and Anderson, D. R. 2004. Multimodel inference understanding AIC and BIC in model selection. *Sociological methods and research*, 33(2):261–304. doi: 10.1177/0049124104268644.

Calderhead, B. and Girolami, M. 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, 53(12):4028–4045. doi: http://dx.doi.org/10.1016/j.csda.2009.07.025.

Cardinali, C., Pezzulli, S., and Andersson, E. 2004. Influence-matrix diagnostic of a data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 130(603):2767–2786. doi: 10.1256/qj.03.205.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. 2017. Stan : A probabilistic programming language. *Journal of Statistical Software*, 76(1). doi: 10.18637/jss.v076.i01.

Cartwright, N. 1983. *How the Laws of Physics Lie. How the Laws of Physics Lie*. New York: Oxford University Press.

Cirpka, O. A. and Nowak, W. 2004. First-order variance of travel time in nonstationary formations. *Water Resources Research*, 40(3). doi: 10.1029/2003WR002851.

Cirpka, O. A., de Barros, F. P. J., Chiogna, G., Rolle, M., and Nowak, W. Stochastic flux-related analysis of transverse mixing in two-dimensional heterogeneous porous media. *Water Resources Research*, 47(6). doi: 10.1029/2010WR010279.

Claeskens, G. 2016. Statistical model choice. *Annual Review of Statistics and Its Application*, 3:233–256.

Claeskens, G., Hjort, N. L., et al. 2008. Model selection and model averaging. *Cambridge Books*.

Clark, J. S. and Gelfand, A. E. 2006. *Hierarchical Modelling for the Environmental Sciences: Statistical methods and applications. Hierarchical Modelling for the Environmental Sciences: Statistical methods and applications*. Oxford University Press, ISBN: 0191513849.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E. 2008. Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44(12).

Clark, M. P., Kavetski, D., and Fenicia, F. 2011. Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, 47 (9). doi: 10.1029/2010WR009827.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., et al. 2015. A unified approach for process-based hydrologic modeling: 2. model implementation and case studies. *Water Resources Research*, 51(4):2515–2542.

Clarke, B. 2003. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research*, 4(Oct):683–712.

Clyde, M. A. and Iversen, E. S. 2013. Bayesian model averaging in the M-open framework. *Bayesian theory and applications*.

Cotter, S. L., Roberts, G. O., Stuart, A. M., and White, D. 2013. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical science*, 28(3):424–446.

Cucker, F. and Smale, S. 2002. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49.

de Barros, F. P. and Nowak, W. 2010. On the link between contaminant source release conditions and plume prediction uncertainty. *Journal of contaminant hydrology*, 116(1-4):24–34.

DeGroot, M. H. 2005. *Optimal statistical decisions*. volume 82. *Optimal statistical decisions*. John Wiley & Sons.

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J. 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrol. Earth Syst. Sci.*, 17(10):4209–4225. doi: 10.5194/hess-17-4209-2013.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Diks, C. G. and Vrugt, J. A. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment*, 24(6):809–820.

Domingos, P. Bayesian averaging of classifiers and the overfitting problem. In *ICML*, pages 223–230, 2000.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97.

Du, J. 2016. The "weight" of models and complexity. *Complexity*, 21(3):21–35. doi: 10.1002/cplx.21612.

Earl, D. J. and Deem, M. W. 2005. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. pages 1–26. doi: 10.1214/aos/1176344552.

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the american statistical association*, 78(382):316–331. doi: 10.1080/01621459.1983.10477973.

Efron, B. 1986. How biased is the apparent error rate of a prediction rule? *Journal of the american statistical association*, 81(394):461–470. doi: 10.2307/2289236.

Efron, B. and Tibshirani, R. J. 1994. *An introduction to the bootstrap. An introduction to the bootstrap*. CRC press, ISBN: 0412042312.

Elliott, L. P. and Brook, B. W. 2007. Revisiting chamberlin: multiple working hypotheses for the 21st century. *AIBS Bulletin*, 57(7):608–614.

Elshall, A. S. and Tsai, F. T.-C. 2014. Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. *Journal of Hydrology*, 517:105 – 119. ISSN 0022-1694. doi: https://doi.org/10.1016/j.jhydrol.2014.05.027.

Faust, J., Gilchrist, S., Wright, J. H., and Zakrajšsek, E. 2013. Credit spreads as predictors of real-time economic activity: a Bayesian model-averaging approach. *Review of Economics and Statistics*, 95(5):1501–1519. doi: 10.1162/REST-a-00376.

Fenicia, F., Kavetski, D., Savenije, H. H. G., and Pfister, L. 2016. From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. *Water Resources Research*. doi: 10.1002/2015WR017398.

Ferré, T. P. 2017. Revisiting the relationship between data, models, and decision-making. *Groundwater*, 55(5):604–614.

Fisher, R. 1922. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. 2013. emcee: the MCMC hammer. *Publications of the Astronomical Society of the Pacific*, 125 (925):306.

Freeze, R. A. and Harlan, R. 1969. Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of Hydrology*, 9(3):237–258.

Friedman, J., Hastie, T., and Tibshirani, R. 2001. *The elements of statistical learning*. In *Springer Ser. Stat.*, volume 1. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, ISBN: 9780387216065.

Friel, N. and Pettitt, A. N. 2008. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607. doi: 10.1111/j.1467-9868.2007.00650.x.

Friel, N. and Wyse, J. 2012. Estimating the evidence – a review. *Statistica Neerlandica*, 66(3):288–308. doi: 10.1111/j.1467-9574.2011.00515.x.

Friel, N., Hurn, M., and Wyse, J. 2013. Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24(5):709–723. doi: 10.1007/s11222-013-9397-1.

Friel, N., McKeone, J. P., Oates, C. J., and Pettitt, A. N. 2016. Investigation of the widely applicable Bayesian information criterion. *Statistics and Computing*, pages 1–12. doi: 10.1007/s11222-016-9657-y.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. 2017. Visualization in Bayesian workflow. *arXiv preprint arXiv:1709.01449*.

Garthwaite, P. H. and Mubwandarikwa, E. 2010. Selection of weights for weighted model averaging. *Australian & New Zealand Journal of Statistics*, 52(4):363–382.

Geisser, S. and Eddy, W. F. 1979. A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160. doi: 10.1080/01621459.1979.10481632.

Gelfand, A. E. and Dey, D. K. 1994. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Gell-Mann, M. 1995. What is complexity? remarks on simplicity and complexity by the nobel prize-winning author of the quark and the jaguar. *Complexity*, 1 (1):16–19. doi: 10.1002/cplx.6130010105.

Gelman, A. and Meng, X.-L. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 13(2):163–185.

Gelman, A. and Rubin, D. B. 1992. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472. doi: 10.2307/2246093.

Gelman, A. and Shalizi, C. R. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1): 8–38. doi: 10.1111/j.2044-8317.2011.02037.x.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 2004. *Bayesian data analysis*. volume 2. *Bayesian data analysis*. Chapman and Hall/CRC, ISBN: 9781584883883.

Gelman, A., Hwang, J., and Vehtari, A. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016. doi: 10.1007/s11222-013-9416-2.

Gelman, A. et al. 2008. Objections to Bayesian statistics. *Bayesian Analysis*, 3 (3):445–449.

Gembris, D., Taylor, J. G., and Suter, D. 2007. Evolution of athletic records: statistical effects versus real improvements. *Journal of Applied Statistics*, 34(5): 529–545.

George, E. I. et al. Dilution priors: Compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics, 2010.

Ghahramani, Z. 2013. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1984). doi: 10.1098/rsta.2011.0553.

Gillies, D. 2012. *Philosophical theories of probability. Philosophical theories of probability*. Routledge.

Giudice, D. D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., and Rieckermann, J. 2013. Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. *Hydrology and Earth System Sciences*, 17(10): 4209–4225.

Gneiting, T. and Raftery, A. E. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Golder, J., Joelson, M., Neel, M.-C., and Di Pietro, L. 2014. A time fractional model to represent rainfall process. *Water Science and Engineering*, 7(1):32–40. doi: http://dx.doi.org/10.3882/j.issn.1674-2370.2014.01.004.

Good, I. J. 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114.

Granger, C. W. J. and Ramanathan, R. 1984. Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204. doi: 10.1002/for.3980030207.

Green, P. J. 1995. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732. doi: 10.1093/biomet/82.4.711.

Grünwald, P. 2000. Model selection based on minimum description length. *Journal of mathematical psychology*, 44(1):133–152. doi: http://dx.doi.org/10.1006/jmps.1999.1280.

Grünwald, P. and Vitányi, P. B. 2003. Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, 12(4):497–529. doi: 10.1023/A:1025011119492.

Gupta, H. V. and Razavi, S. 2018. Revisiting the basis of sensitivity analysis for dynamical earth system models. *Water Resources Research*, 0(0):in press. doi: 10.1029/2018WR022668.

Guthke, A. 2017. Defensible model complexity: A call for data-based and goal-oriented model choice. *Groundwater*, 55(5):646–650. doi: 10.1111/gwat.12554.

Guyon, I., Saffari, A., Dror, G., and Cawley, G. 2010. Model selection: Beyond the Bayesian/frequentist divide. *J. Mach. Learn. Res.*, 11:61–87.

Hammersley, J. and Handscomb, D. 1964. *Monte Carlo Methods*. In *Monographs on Applied Probability and Statistics. Monte Carlo Methods*. Methuen & Co., London, and John Wiley & Sons, New York.

Hannan, E. J. and Quinn, B. G. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195.

Hansen, M. H. and Yu, B. 2001. Model selection and the principle of minimum description length. *Journal of the american statistical association*, 96(454):746–774. doi: 10.1198/016214501753168398.

Harding, S. 1975. *Essays on the Duhem-Quine thesis*. In *Can theories be refuted?*, volume 81. *Essays on the Duhem-Quine thesis*. Springer Science & Business Media.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. 1999. Bayesian model averaging: A tutorial. *Statistical science*, 14(4):382–401. doi: 10.1214/ss/1009212519.

Hoffman, M. D. and Gelman, A. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Höge, M., Wöhling, T., and Nowak, W. 2018. A primer for model selection: The decisive role of model complexity. *Water Resources Research*, 54(3):1688–1715. doi: 10.1002/2017WR021902.

Höge, M., Guthke, A., and Nowak, W. 2019. The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology*, 572:96–107. doi: 10.1016/j.jhydrol.2019.01.072.

Hooten, M. B. and Hobbs, N. T. 2015. A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1):3–28. doi: DOI10.1890/14-0661.1.sm.

Hsu, K.-l., Gupta, H. V., and Sorooshian, S. 1995. Artificial neural network modeling of the rainfall-runoff process. *Water resources research*, 31(10):2517–2530.

Huelsenbeck, J. P. and Ronquist, F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Hurvich, C. M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307. doi: 10.2307/2336663.

Hutter, M. 2007. On universal prediction and Bayesian confirmation. *arXiv preprint arXiv:0709.1516*.

Illman, W. A., Zhu, J., Craig, A. J., and Yin, D. 2010. Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study. *Water Resources Research*, 46(4). doi: ArtnW0450210.1029/2009wr007745.

Janson, L., Fithian, W., and Hastie, T. J. 2015. Effective degrees of freedom: a flawed metaphor. *Biometrika*. doi: 10.1093/biomet/asv019.

Juárez-Ruiz, E., Cortés-Maldonado, R., and Pérez-Rodríguez, F. 2016. Relationship between the inverses of a matrix and a submatrix. *Computación y Sistemas*, 20(2):251–262.

Kadane, J. B. and Lazar, N. A. 2004. Methods and criteria for model selection. *Journal of the American Statistical Association*, 99(465):279–290. doi: 10.1198/016214504000000269.

Kashyap, R. L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 4(2):99–104. doi: 10.1109/TPAMI.1982.4767213.

Kass, R. E. and Raftery, A. E. 1995. Bayes factors. *Journal of the american statistical association*, 90(430):773–795. doi: 10.1080/01621459.1995.10476572.

Kavetski, D., Kuczera, G., and Franks, S. W. 2006. Bayesian analysis of input uncertainty in hydrological modeling: 2. application. *Water Resources Research*, 42(3). doi: 10.1029/2005WR004376.

Kim, H.-C. and Ghahramani, Z. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, pages 619–627, 2012.

Klenke, A. 2013. *A comprehensive course*. In *Probability theory. A comprehensive course*. Springer Science & Business Media.

Klesov, O. 2014. *Limit theorems for multi-indexed sums of random variables*. volume 71. *Limit theorems for multi-indexed sums of random variables*. Springer.

Klir, G. and Yuan, B. 1995. *Fuzzy sets and fuzzy logic*. volume 4. *Fuzzy sets and fuzzy logic*. Prentice hall New Jersey.

Lanterman, A. D. 2001. Schwarz, wallace, and rissanen: Intertwining themes in theories of model selection. *International Statistical Review / Revue Internationale de Statistique*, 69(2):185–212. doi: 10.2307/1403813.

Lartillot, N. and Philippe, H. 2006. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207. doi: 10.1080/10635150500433722.

Le, T., Clarke, B., et al. 2017. A Bayes interpretation of stacking for M-complete and M-open settings. *Bayesian Analysis*, 12(3):807–829. doi: 10.1214/16-BA1023.

Leeb, H. and Pötscher, B. M. 2009. *Model Selection*. In Mikosch, T., Kreiss, J.-P., Davis, R. A., and Andersen, T. G., editors, *Handbook of Financial Time Series*, pages 889–925. Springer Berlin Heidelberg. doi: 10.1007/978-3-540-71297-839.

Leube, P. C., de Barros, F. P. J., Nowak, W., and Rajagopal, R. 2013. Towards optimal allocation of computer resources: Trade-offs between uncertainty quantification, discretization and model reduction. *Environmental Modelling & Software*, 50:97–107. doi: http://dx.doi.org/10.1016/j.envsoft.2013.08.008.

Li, Y. F., Ng, S. H., Xie, M., and Goh, T. N. 2010. A systematic comparison of metamodeling techniques for simulation optimization in decision support systems. *Applied Soft Computing*, 10(4):1257–1273. doi: http://dx.doi.org/10.1016/j.asoc.2009.11.034.

Liu, P. G., Elshall, A. S., Ye, M., Beerli, P., Zeng, X. K., Lu, D., and Tao, Y. Z. 2016. Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resources Research*, 52(2):734–758. doi: 10.1002/2014wr016718.

Liu, Y. and Gupta, H. V. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43(7). doi: 10.1029/2006WR005756.

Lu, D., Ye, M., and Neuman, S. P. 2011. Dependence of Bayesian model selection criteria and fisher information matrix on sample size. *Mathematical Geosciences*, 43(8):971–993. doi: 10.1007/s11004-011-9359-0.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. 2000. Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.

MacKay, D. J. 1992. Bayesian interpolation. *Neural computation*, 4(3):415–447. doi: DOI10.1162/neco.1992.4.3.415.

Mallick, H. and Yi, N. 2013. Bayesian methods for high dimensional linear models. *Journal of biometrics & biostatistics*, 1:005. doi: 10.4172/2155-6180.S1-005.

Mallows, C. 1973. Some comments on Cp. *Technometrics*, 15(4):661–675. doi: 10.2307/1267380.

Marconato, A., Schoukens, M., Rolain, Y., and Schoukens, J. Study of the effective number of parameters in nonlinear identification benchmarks. In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 4308–4313, 2013.

Mattingly, H. H., Transtrum, M. K., Abbott, M. C., and Machta, B. B. 2018. Maximizing the information learned from finite data selects a simple model. *Proceedings of the National Academy of Sciences*, 115(8):1760–1765.

McKay, M. D., Beckman, R. J., and Conover, W. J. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245. ISSN 00401706. URL `http://www.jstor.org/stable/1268522`.

McQuarrie, A. D. 1999. A small-sample correction for the schwarz sic model selection criterion. *Statistics & Probability Letters*, 44(1):79–86. doi: https://doi.org/10.1016/S0167-7152(98)00294-6.

McQuarrie, A. D. and Tsai, C.-L. 1998. *Regression and time series model selection. Regression and time series model selection.* World Scientific, ISBN: 978-9810232429.

Mekonnen, B. A., Nazemi, A., Mazurek, K. A., Elshorbagy, A., and Putz, G. 2015. Hybrid modelling approach to prairie hydrology: fusing data-driven and process-based hydrological models. *Hydrological sciences journal*, 60(9):1473–1489.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H. 2015. Are we unnecessarily constraining the agility of complex process-based models? *Water Resources Research*, 51(1):716–728. doi: 10.1002/2014WR015820.

Meyer, R. Deviance information criterion (DIC). In *Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd, ISBN: 9781118445112*, 2014.

Minka, T. P. 2002. Bayesian model averaging is not model combination. *Available electronically at http://www. stat. cmu. edu/minka/papers/bma. html.*

Mononen, T. 2015. A case study of the widely applicable Bayesian information criterion and its optimality. *Statistics and Computing*, 25(5):929–940. doi: 10.1007/s11222-014-9463-3.

Montanari, A. 2007. What do we mean by 'uncertainty'? the need for a consistent wording about uncertainty assessment in hydrology. *Hydrological Processes: An International Journal*, 21(6):841–845.

Montanari, A. and Koutsoyiannis, D. 2012. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research*, 48(9).

Monteith, K., Carroll, J. L., Seppi, K., and Martinez, T. Turning Bayesian model averaging into Bayesian model combination. In *The 2011 International Joint Conference on Neural Networks*, July 2011. doi: 10.1109/IJCNN.2011.6033566.

Moody, J., Hanson, S., and Lippmann, R. 1992. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. *Advances in neural information processing systems*, 4:847–854. doi: 10.1.1.28.295.

Moore, R. E. 1979. *Methods and applications of interval analysis.* volume 2. *Methods and applications of interval analysis.* SIAM.

Myung, I. J., Balasubramanian, V., and Pitt, M. A. 2000. Counting probability distributions: differential geometry and model selection. *Proc Natl Acad Sci U S A*, 97(21):11170–5. doi: 10.1073/pnas.170283897.

Myung, J. I., Navarro, D. J., and Pitt, M. A. 2006. Model selection by normalized maximum likelihood. *Journal of mathematical psychology*, 50(2):167–179. doi: https://doi.org/10.1016/j.jmp.2005.06.008.

Neal, R. M. 1996. *Bayesian learning for neural networks*. In *Lect. Notes in Stat.*, volume 1. *Bayesian learning for neural networks*. Springer Science & Business Media.

Nearing, G. and Gupta, H. 2018. Ensembles vs. information theory: Supporting science under uncertainty. *Frontiers of Earth Science*, in revision.

Nearing, G. S. and Gupta, H. V. 2015. The quantity and quality of information in hydrologic models. *Water Resources Research*, 51(1):524–538. doi: 10.1002/2014WR015895.

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., and Weijs, S. V. 2016. A philosophical basis for hydrological uncertainty. *Hydrol. Sci. J.*, 61(9):1666–1678. doi: 10.1080/02626667.2016.1183009.

Neuman, S. P. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment*, 17(5):291–305. doi: 10.1007/s00477-003-0151-7.

Newton, M. A. and Raftery, A. E. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48.

Nowak, W. and Guthke, A. 2016. Entropy-based experimental design for optimal model discrimination in the geosciences. *Entropy*, 18(11):409. doi: 10.3390/e18110409.

Nowak, W., Schwede, R. L., Cirpka, O. A., and Neuweiler, I. 2008. Probability density functions of hydraulic head and velocity in three-dimensional heterogeneous porous media. *Water Resources Research*, 44(8). doi: 10.1029/2007WR006383.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. 2006. *Uncertain judgements: eliciting experts' probabilities. Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

Omlin, M. and Reichert, P. 1999. A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling*, 115(1):45–59.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M. 2015. Does model performance improve with complexity? a case study with three hydrological models. *Journal of Hydrology*, 523:147–159. doi: https://doi.org/10.1016/j.jhydrol.2015.01.044.

Oxford. Definition of: "model". In *OxfordDictionaries.com*. Oxford University Press. URL `https://en.oxforddictionaries.com/definition/model`. [Accessed Dec. 31, 2018].

Pande, S., McKee, M., and Bastidas, L. A. 2009. Complexity-based robust hydrologic prediction. *Water Resources Research*, 45(10):3945–4004. doi: 10.1029/2008WR007524.

Pande, S., Arkesteijn, L., Savenije, H., and Bastidas, L. A. 2015. Hydrological model parameter dimensionality is a weak measure of prediction uncertainty. *Hydrol. Earth Syst. Sci. Discuss.*, 12(4):3945–4004. doi: 10.5194/hessd-12-3945-2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Piironen, J. and Vehtari, A. 2017. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735. doi: 10.1007/s11222-016-9649-y.

Plummer, M. et al. Jags: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria, 2003.

Poeter, E. and Anderson, D. 2005. Multimodel ranking and inference in ground water modeling. *Ground Water*, 43(4):597–605. doi: 10.1111/j.1745-6584.2005.0061.x.

Popper, K. 2005. *The logic of scientific discovery. The logic of scientific discovery.* Routledge.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174. doi: 10.1175/MWR2906.1.

Rathmanner, S. and Hutter, M. 2011. A philosophical treatise of universal induction. *Entropy*, 13(6):1076–1136.

Ray, D. and Hesthaven, J. S. 2018. An artificial neural network as a troubled-cell indicator. *Journal of computational physics*, 367:166–191.

Refsgaard, J. C. and Abbott, M. B. 1996. *The Role of Distributed Hydrological Modelling in Water Resources Management.* In Abbott, M. B. and Refsgaard, J. C., editors, *Distributed Hydrological Modelling*, volume 22 of *Water Science and Technology Library*, chapter 1, pages 1–16. Springer Netherlands.

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., and Troldborg, L. 2012. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Adv. Water Resour.*, 36:36–50. doi: 10.1016/j.advwatres.2011.04.006.

Reichert, P. and Mieleitner, J. 2009. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research*, 45(10).

Reichert, P., Langhans, S. D., Lienert, J., and Schuwirth, N. 2015. The conceptual foundation of environmental decision support. *Journal of environmental management*, 154:316–332.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W. 2010. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, 46(5). doi: ArtnW0552110.1029/2009wr008328.

Rinderknecht, S. L., Borsuk, M. E., and Reichert, P. 2012. Bridging uncertain and ambiguous knowledge with imprecise probabilities. *Environ. Model. Softw.*, 36:122–130. ISSN 1364-8152. doi: 10.1016/j.envsoft.2011.07.022.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica*, 14(5): 465–471.

Rissanen, J. 1987. Stochastic complexity and the mdl principle. *Econometric Reviews*, 6(1):85–102. doi: 10.1080/07474938708800126.

Rissanen, J. J. 1996. Fisher information and stochastic complexity. *Ieee Transactions on Information Theory*, 42(1):40–47. doi: Doi10.1109/18.481776.

Rubin, D. B. 1981. The Bayesian bootstrap. *The annals of statistics*, pages 130–134.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55.

Samaniego, L., Kumar, R., and Attinger, S. 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46(5). doi: 10.1029/2008WR007327.

Sanderson, B. M., Knutti, R., and Caldwell, P. 2015. Addressing interdependency in a multimodel ensemble by interpolation of model properties. *Journal of Climate*, 28(13):5150–5170.

Schlaifer, R. and Raiffa, H. 1961. *Applied statistical decision theory. Applied statistical decision theory*. Havard University.

Schöniger, A. *Bayesian assessment of conceptual uncertainty in hydrosystem modeling*. PhD thesis, Doctoral dissertation, Universität Tübingen, Tübingen, Germany, 2016.

Schöniger, A., Wöhling, T., Samaniego, L., and Nowak, W. 2014. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research*, 50(12):9484–9513. doi: 10.1002/2014WR016062.

Schöniger, A., Illman, W. A., Wöhling, T., and Nowak, W. 2015a. Finding the right balance between groundwater model complexity and experimental effort via Bayesian model selection. *Journal of Hydrology*, 531:96–110. doi: 10.1016/j.jhydrol.2015.07.047.

Schöniger, A., Wöhling, T., and Nowak, W. 2015b. A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. *Water Resources Research*, 51(9):7524–7546. doi: 10.1002/2015WR016918.

Schoups, G. and Vrugt, J. A. 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-gaussian errors. *Water Resources Research*, 46(10). doi: 10.1029/2009WR008933.

Schoups, G., van de Giesen, N. C., and Savenije, H. H. G. 2008. Model complexity control for hydrologic prediction. *Water Resources Research*, 44(12):1944–1973. doi: 10.1029/2008WR006836.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. doi: doi:10.1214/aos/1176344136.

Schweppe, F. C. 1973. *Uncertain dynamic systems. Uncertain dynamic systems.* Prentice Hall.

Shannon, C. E. 1948. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shibata, R. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The annals of statistics*, 8(1):147–164.

Shibata, R. 1986. Consistency of model selection and parameter estimation. *Journal of Applied Probability*, 23:127–141. doi: 10.2307/3214348.

Shiffrin, R. M., Chandramouli, S. H., and Grünwald, P. D. 2016. Bayes factors, relations to minimum description length, and overlapping model classes. *Journal of mathematical psychology*, 72:56–77. doi: http://dx.doi.org/10.1016/j.jmp.2015.11.002.

Simsekli, U., Badeau, R., Richard, G., and Cemgil, A. T. Stochastic thermodynamic integration: efficient Bayesian model selection via stochastic gradient MCMC. In *41st International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

Sinsbeck, M. and Tartakovsky, D. M. 2015. Impact of data assimilation on cost-accuracy tradeoff in multifidelity models. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):954–968.

Skilling, J. Nested sampling. In *AIP Conference Proceedings*, volume 735, pages 395–405. AIP, 2004.

Solomonoff, R. J. 1964. A formal theory of inductive inference. part i. *Information and control*, 7(1):1–22.

Sorooshian, S. and Dracup, J. A. 1980. Stochastic parameter estimation procedures for hydrologie rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resources Research*, 16(2):430–442. doi: 10.1029/WR016i002p00430.

Spiegelhalter, D. J., Best, N. G., Carlin, B. R., and van der Linde, A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64(4):583–616. doi: Doi10.1111/1467-9868.00353.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493. doi: 10.1111/rssb.12062.

Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47.

Tao, T. 2012. E pluribus unum: from complexity, universality. *Daedalus*, 141(3): 23–34.

Tarantola, A. 2005. *Inverse problem theory and methods for model parameter estimation. Inverse problem theory and methods for model parameter estimation.* SIAM.

Tarantola, A. 2006. Popper, Bayes and the inverse problem. *Nature physics*, 2(8): 492–494.

Tibshirani, R. J. 2014. Degrees of freedom and model search. *arXiv preprint arXiv:1402.1920*.

Tribus, M. 1961. Information theory as the basis for thermostatics and thermodynamics. *Journal of Applied Mechanics*, 28(1):1–8.

Trotta, R. 2008. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104. doi: 10.1080/00107510802066753.

Vaiter, S., Golbabaee, M., Fadili, J., and Peyré, G. 2015. Model selection with low complexity priors. *Information and Inference*. doi: 10.1093/imaiai/iav005.

van der Linde, A. 2012. A Bayesian view of model complexity. *Statistica Neerlandica*, 66(3):253–271. doi: 10.1111/j.1467-9574.2011.00518.x.

Vandekerckhove, J., Matzke, D., and Wagenmakers, E.-J. Model comparison and the principle of parsimony. In Townsend, J. T. and Busemeyer, J. R. B., editors, *The Oxford Handbook of Computational and Mathematical Psychology*, pages 300–319. Oxford University Press, ISBN: 9780199957996, doi: 10.1093/oxfordhb/9780199957996.013.14, 2015.

VanderPlas, J. Frequentism and Bayesianism: A Python-driven primer. In van der Walt, S. and Bergstra, J., editors, *13th PYTHON IN SCIENCE CONF. (SCIPY 2014)*, pages 91–99, 2014.

Vanpaemel, W. Measuring model complexity with the prior predictive. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1919–1927, 2009. Red Hook, NY: Curran Associates Inc.

Vehtari, A. and Ojanen, J. 2012. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228. doi: 10.1214/12-SS102.

Vehtari, A., Gelman, A., and Gabry, J. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27 (5):1413–1432.

Vitoratou, S. and Ntzoufras, I. 2017. Thermodynamic Bayesian model comparison. *Statistics and Computing*, 27(5):1165–1180.

von Gunten, D., Wöhling, T., Haslauer, C., Merchán, D., Causapé, J., and Cirpka, O. A. 2014. Efficient calibration of a distributed pde-based hydrological model using grid coarsening. *Journal of Hydrology*, 519, Part D:3290–3304. doi: http://dx.doi.org/10.1016/j.jhydrol.2014.10.025.

Vousden, W. D., Farr, W. M., and Mandel, I. 2016. Dynamic temperature selection for parallel tempering in Markov Chain Monte Carlo simulations. *Monthly Notices of the Royal Astronomical Society*, 455(2):1919–1937. doi: 10.1093/mnras/stv2422.

Vrieze, S. I. 2012. Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2):228–243. doi: 10.1037/a0027127.

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., and Robinson, B. A. 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov Chain Monte Carlo simulation. *Water Resources Research*, 44(12).

Vrugt, J. A., Ter Braak, C., Diks, C., Robinson, B. A., Hyman, J. M., and Higdon, D. 2009. Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290.

Wagener, T., Reed, P., van Werkhoven, K., Tang, Y., and Zhang, Z. 2009. Advances in the identification and evaluation of complex environmental systems models. *Journal of Hydroinformatics*, 11(3-4):266–281.

Watanabe, S. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594.

Watanabe, S. 2013. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897.

Weijs, S. V., Van Nooijen, R., and Van De Giesen, N. 2010. Kullback–leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138(9):3387–3399.

Wit, E., Heuvel, E. v. d., and Romeijn, J.-W. 2012. 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236. doi: 10.1111/j.1467-9574.2012.00530.x.

Wöhling, T., Schöniger, A., Gayler, S., and Nowak, W. 2015. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. *Water Resources Research*, 51(4):2825–2846. doi: 10.1002/2014WR016292.

Wolpert, D. H. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950. doi: 10.1093/biomet/92.4.937.

Yang, Y. and Dunson, D. B. 2013. Sequential Markov Chain Monte Carlo. *arXiv preprint arXiv:1308.3861*.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al. 2017. Using stacking to average Bayesian predictive distributions. *https://projecteuclid.org/euclid.ba/1516093227*. doi: 10.1214/17-BA1091.

Ye, J. 1998. On measuring and correcting the effects of data mining and model selection. *Journal of the american statistical association*, 93(441):120–131. doi: 10.2307/2669609.

Ye, M., Neuman, S. P., and Meyer, P. D. 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, 40(5).

Ye, M., Meyer, P. D., and Neuman, S. P. 2008. On model selection criteria in multimodel analysis. *Water Resources Research*, 44(3). doi: ArtnW0342810. 1029/2008wr006803.

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M., and Reeves, D. M. 2010. A model-averaging method for assessing groundwater conceptual model uncertainty. *Groundwater*, 48(5):716–728.

Zadeh, L. 1965. Fuzzy sets. *Information and Control*, 8(3):338 – 353. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(65)90241-X.

Zadeh, L. A. 1999. Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems*, 100(1):9–34.

Zhang, X. and Zhao, K. 2012. Bayesian neural networks for uncertainty analysis of hydrologic modeling: A comparison of two schemes. *Water Resources Management*, 26(8):2365–2382. doi: 10.1007/s11269-012-0021-5.

Zou, H., Hastie, T., and Tibshirani, R. 2007. On the "degrees of freedom" of the LASSO. *Ann. Statist.*, 35(5):2173–2192. doi: 10.1214/009053607000000127.