Diagnosis of Model Errors With a Sliding Time-Window Bayesian Analysis

Han-Fang Hsueh[1,2] [ID], **Anneli Guthke[1,3]** [ID], **Thomas Wöhling[4,5]** [ID], and **Wolfgang Nowak[1]** [ID]

[1]Department of Stochastic Simulation and Safety Research for Hydrosystems (IWS/LS3), University of Stuttgart, Stuttgart, Germany, [2]Center for Applied Geoscience, University of Tübingen, Tübingen, Germany, [3]Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart, Germany, [4]Department of Hydrology, Technical University of Dresden, Dresden, Germany, [5]Lincoln Environmental Research, Lincoln Agritech, Hamilton, New Zealand

Abstract Deterministic hydrological models with uncertain, but inferred-to-be-time-invariant parameters typically show time-dependent model errors. Such errors can occur if a hydrological process is active in certain time periods in nature, but is not resolved by the model or by its input. Such missing processes could become visible during calibration as time-dependent best-fit values of model parameters. We propose a formal time-windowed Bayesian analysis to diagnose this type of model error, formalizing the question "In which period of the calibration time-series does the model statistically disqualify itself as quasi-true?" Using Bayesian model evidence (BME) as model performance metric, we determine how much the data in time windows of the calibration time-series support or refute the model. Then, we track BME over sliding time windows to obtain a dynamic, time-windowed BME (tBME) and search for sudden decreases that indicate an onset of model error. tBME also allows us to perform a formal, sliding likelihood-ratio test of the model against the data. Our proposed approach is designed to detect error occurrence on various temporal scales, which is especially useful in hydrological modeling. We illustrate this by applying our proposed method to soil moisture modeling. We test tBME as model error indicator on several synthetic and real-world test cases that we designed to vary in error sources (structure and input) and error time scales. Results prove the successful detection errors in dynamic models. Moreover, the time sequence of posterior parameter distributions helps to investigate the reasons for model error and provide guidance for model improvement.

1. Introduction

Water resources research and management is increasingly relying on mathematical models to gain a deeper system understanding, to predict a system's current state (now-casting), and to predict a system's state under future conditions (forecasting). Unfortunately, from a philosophical point of view, all models are wrong because they are simplifications of the real system (Box, 1976). Previous research has studied the types of errors in hydrologic models, including input error from measurements (Ajami et al., 2007; Kavetski et al., 2006; McMillan et al., 2018; Mustafa et al., 2018), parameter errors from uncertain estimated model parameter values (Bastola et al., 2008; Dobler et al., 2012; Remondi et al., 2018; Teweldebrhan et al., 2018), and model-structural error, which stems from the wrong setup of a model. We will use the term "model error" to refer to both model-structural error and model input error. In the following, assuming that parameter errors are treated by assigning (wide enough) parameter distributions to the model structure. Model errors generally arise when the dominant processes are not comprehensively and correctly represented in a model. Examples of misspecified processes or missing processes in hydrological models are given by Gupta et al. (2012). Model uncertainties and errors typically exist in model inputs (forcings; Del Giudice et al., 2013, 2015; Kuczera et al., 2006), model building (Clark et al., 2008), and observations of model output (Renard et al., 2010). Biased or non-representative input data will unevitably lead to biased model output if mass conservation is enforced by the model structure. It is important to investigate the occurrence of model errors for several reasons: (a) Model errors lead to biased predictions (Ajami et al., 2007; Brynjarsdóttir & O'Hagan, 2014) and hence compromise the usefulness of a model for practically all modeling purposes (process identification, nowcasting, forcasting) and subsequent management decisions (Sargsyan et al., 2019; Steinschneider et al., 2015). (b) Model errors lead to biased parameter identification through calibration or Bayesian parameter inference (Xu & Valocchi, 2015) or cause non-identifiability of parameters (Menberg et al., 2019; Reichert & Mieleitner, 2009), because parameters are optimized in the wrong context; they are adjusted regardless of their physically feasible ranges to obtain a satisfying agreement to observed data despite the model error (Oreskes et al., 1994). (c) Given model errors, uncertainty analysis

within a conventional Bayesian framework yields misleading results, because Bayesian theory assumes an error-free ("true") model (Fenicia et al., 2018; Gelman et al., 2013; Box & Tiao, 1992; Xu et al., 2017). (d) With the investigation and resolution of model errors, we have the chance to get the right answers for the right reasons (Kirchner, 2006) and improve the scientific understanding of a system (Andréassian et al., 2009).

Model errors pose a problem, if we are interested in inferring realistic parameter values that shall relate to physical properties of the hydrological system, or if we wish to identify dominant physical processes, or if we aim to make predictions by extrapolating a physically-based model beyond calibration conditions. When parameters compensate for model errors, they lose their physical meaning (Reichert & Schuwirth, 2012; Xu et al., 2017). Such errors can also show more subtle, for example, through the fact that dynamic parameters or a nonstationary model status improve model performance (Getahun & Van Lanen, 2015; Morán-Tejeda et al., 2015; Pianosi & Wagener, 2016; Wagner et al., 2016). Examples of processes that often cause time-dependent model structures in hydrological systems are soil cracking-shrinking and freezing-thawing cycles, complex vegetation dynamics, or transient changes of river morphology. Model errors can also be caused by neglecting major processes (e.g., macro-pore flow) or by the application of ill-suited model concepts. Examples of input errors are inaccurate meteorological data, for example, when using radar-derived rainfall or when inter- or extrapolating from single rain gauge to a different location at a different scale.

Model errors are currently addressed by methods that typically follow one of two primary directions: error detection or error mitigation. To detect error occurrence, the temporal change of model performance is often used as an indicator. Klemeš (1986) presented a generalization of the routine split-sample test. Their study shows varied model performance in the calibration period, which is associated with varied model conditions. Following this concept, Coron et al. (2012) proposed the generalized split-sample test that uses a sliding window to calibrate models with subsets of the time series data and to study model deficiencies by analyzing the parameter transferability. Motavita et al. (2019) proposed the comprehensive differential split-sample test (CDSST) to study (non-) stationarity of calibrated parameters. This research even suggests that, for a model with errors, the choice of hydrological conditions in calibration is more important for model reliability than the length of the calibration dataset. Previous studies have proposed to use parameter nonstationarity as a hint to diagnose model errors (Wagener et al., 2001, 2003; Westra et al., 2014; Wilby, 2005). Pathiraja et al. (2016) allow parameters to vary in time and use data assimilation to detect temporal patterns. While the mentioned approaches rely on common model evaluation metrics, Ruddell et al. (2019) propose to diagnose structural error by quantifying the trade-off between functional and predictive performance with information-theoretic measures.

Another approach to detect structural errors is to compare the performance of competing model candidates in hypothesis-testing approaches with modular modeling frameworks (e.g., Clark et al., 2015, 2008; Fenicia et al., 2011; Knoben et al., 2020) or via Bayesian Model Averaging (Lee et al., 2020; Wöhling et al., 2015). As a specific example, Wöhling et al. (2015) use Bayesian Model Averaging (BMA, Hoeting et al., 1999) to evaluate the worth of different calibration data types. The underlying assumption is that structural errors show by a significant decrease in prediction quality of any specific data type when calibrated on a different data type—parameter compensation is preventing the model from finding a robust best-fit parameter set for all relevant data types. If the preferred model changes depending on the chosen calibration dataset, this is an indication of varying expression of model-structural errors with varied calibration strategy. A review of BMA in a more general context is given by Höge et al. (2019). Del Giudice et al. (2015) assess the decrease of model-structural errors with increasing model complexity and draw conclusions about the relative importance of model-structural errors versus input errors for the quality of model predictions.

To mitigate the impact of model errors on model performance, the majority of approaches treat the apparent bias in model output, that is, these approaches weaken the commonly violated assumption of uncorrelated errors in the likelihood function/objective function during calibration. For example, Ammann et al. (2019) investigate various error models and report that including non-stationary autocorrelation improves model performance. In contrast, however, Barkle et al. (2011) reported unacceptably large prediction errors when using the $AR - 1$ model in the likelihood formation. Smith et al. (2015) summarize the performance of eight residual-error models and report significant change in the posterior mode of their key model parameter as a function of the error model. Generalized likelihood uncertainty estimation (GLUE) (Beven & Binley, 1992) is a variation of the likelihood formulation, which explicitly expresses the autocorrelation in errors and widens the formal likelihood function to subjective limits of acceptability based on expert knowledge (Mirzaei et al., 2015; Ragab et al., 2020). Instead of

correcting the likelihood function, Demissie et al. (2009) propose a data-driven error model to reduce the bias in groundwater model predictions.

Another branch of studies aims to tackle deficits within the model, not only "at the end of the pipe." For example, Xu and Valocchi (2015) infer a Gaussian process error model to improve parameter estimation and model prediction quality. Reichert and Mieleitner (2009) introduce stochastic time-dependent parameters to correct for structural errors within the model. Their goal is to achieve more realistic prediction uncertainty intervals.

In this work, we are focusing on model error detection and localization, because we believe that effective model improvement has to rely on solid error diagnosis as a foundation. This sequential procedure seems scientifically more satisfying than aiming for better model performance without knowing the reasons for the model's misbehavior. We therefore present a method for error detection here, and leave an extension toward error mitigation for future work.

Despite the increasing scientific interest in model errors, existing methods still suffer from major limitations. Current methods for diagnosing model errors require much a-priori, expert knowledge about model errors (Hrachowitz et al., 2014; Westra et al., 2014), trade statistical rigor for easy applicability and/or suffer from numerical inefficiency (Reichert & Mieleitner, 2009). The role and limitations of hypothesis testing for identifying of model errors has been recently discussed by, for example, Nearing et al. (2020), (2016); Neuweiler and Helmig (2017); Beven (2018). Nevertheless, these studies agree that, until now, the hydrological community is lacking a rigorous framework specifically designed for the diagnosis of model errors.

We intend to fill this gap by proposing a time-windowed Bayesian analysis framework to detect and diagnose model errors. We target time-dependent errors in dynamic models or their input, but do not assume any a-priori knowledge about error sources or error patterns. In the spirit of Bayesian hypothesis testing, we identify (localize) periods in the calibration time series that feature statistically significant model error.

We start from the assumption that the model can describe the real world system sufficiently well for most of the time, except when some unresolved processes occasionally occurs. This means that we can claim the model to be quasi-true for specific periods in "time windows," in the sense that it fits the observation data sufficiently well within, for example, one or two days or a week. Only when we increase that time window, we put the model into potential trouble: if the increased time window now contains data impacted by any unresolved process, the calibration goodness-of-fit will deteriorate, and parameter identification will become visibly ambiguous. At that stage, the model parameters are forced to move away from their previous (allegedly) well-calibrated state, in order to compensate for the now occurring model error.

We detect this problematic phase by calibrating the model within a pre-defined shorter time window and moving this time window through the whole data time series (hence the name of our method, "moving time-window analysis"). While moving the time window, we track model performance and search for sudden decreases that would indicate an onset of model error.

We built our framework on the mathematically rigorous grounds of Bayesian probability theory and use Bayesian model evidence (BME) as the model performance metric. BME is perfectly suited to judge a model's performance under parameter and data uncertainty, because it evaluates the model's likelihood to have generated the observed data, averaged over the model's parameter space. We use a brute-force Monte Carlo implementation of BME (Schöniger et al., 2014) as a basis for our proposed routine. This numerical implementation is computationally heavy, but often practically feasible. The additional effort for model error detection is moderate compared to the baseline effort of setting up a predictive ensemble for brute-force Monte Carlo Bayesian updating, because it does not require any additional model runs.

To identify statistically significant drops in performance that are triggered by an onset of model error, we introduce a tBME reference distribution as sampling distribution. We start by asking, what should the tBME value be if the model was true (i.e., error-free)? We randomly sample a realization of model output as modeled observation data to generate a series of theoretical tBME values. We repeat this random sampling procedure many times to produce a random sample of the tBME series that accounts for the full output distribution of the model. This procedure is similar to the bootstrapping analysis on BMS results introduced by Schöniger et al. (2015), with the modeled observation data here being drawn from the model's output distribution instead of being resampled

from noisy observed data. Sampling the distribution of model states to perform "hypothetical calibration" is known from Bayesian pre-posterior analysis (see, e.g., Leube et al., 2012; Nowak & Guthke, 2016). With that tBME sampling distribution, we have a reference for each considered time window. If the tBME value given real data in a specific time window falls outside of the sampled range, it is very unlikely that the model (including its assumptions about parameter and measurement uncertainty) is true, and hence it is very likely that we spotted the onset of model error (i.e., anything that has not been resolved by our model structure, by its inputs, or covered by our uncertainty representation). If, in contrast, the tBME value given a field dataset falls within the reference interval, we conclude that the model fulfills the statistic requirement of a true model at this moment. This model can be considered as a quasi-true model with (statistically) negligible model error. Technically, this is a fully data-driven likelihood ratio test, that is, free of parametric assumption about the distributional shape of model errors.

While the Bayesian approach we are taking required assumptions about measurement uncertainty and parameter uncertainty, these should be understood as part of the model, and not of the proposed method. Our proposed approach to error detection does not presume any model error pattern, and it can handle any choice of formal likelihood function. The method provides an informative and intuitive visualization of model error occurrence on varied temporal scales, simply by changing the size of the sliding window. Using hand-made error in synthetic scenarios, we will further show that this method is able to detect superposed errors, if they occur on different temporal scales (e.g., daily and seasonal patterns). In contrast to most existing methods, our approach follows a data-driven bottom-up perspective: there is no need to make parametric assumptions about the type of model error first and then to formulate and perform a parametric hypothesis test as in Neuweiler and Helmig (2017); rather, we let the observed data speak to identify the onset (and offset) of error in a data-driven hypothesis test. Only in subsequent steps, the modeler may investigate the error period more closely and try to formulate hypotheses about the reasons for model error based on expert intuition.

By detecting model errors with our proposed analysis, we set the grounds for improved system understanding and potential model improvement in further model building. Detection of errors on different temporal scales is especially useful in hydrology, since it helps improve our conceptual understanding in general and because it improves hydrological predictions and their usefulness in all practical aspects. Nevertheless, the application of this method can be generalized to any dynamic system with long time series available for calibration.

In summary, the main contributions of this study are:

1. We develop a data-driven method for model error detection in a statistically rigorous Bayesian framework, with no prior assumptions about error sources or patterns (but assumptions about parameter and measurement uncertainty, possibly also input uncertainty, need to be specified as in any Bayesian context).
2. By introducing a reference range (sampling distribution) for the model performance metric (BME), we formalize the question: "In which periods does the model statistically disqualify itself as quasi-true?" to a data-driven likelihood ratio test.
3. Our proposed approach is designed to detect error occurrence on various temporal scales, with the choice of temporal scales to be investigated left to the modeler. Hence, our framework generalizes well to arbitrary dynamic systems with long calibration time series, but allows for individual adaptation to specific applications and scientific research questions.

The only restriction to our procedure is that the time scale resolved by the available data time-series must be smaller, and the total length of the available data must be larger than the time scale of errors to be diagnosed. To make the assessment observable, the total length of the calibration time series must be longer than the considered time window, so that the signal is analyzed in the context, instead of only a small piece.

The article is structured as follows: Section 2 introduces the proposed method, especially the time-windowed Bayesian analysis and the reference range. Section 3 presents the model and the field data used in our case study, and defines several synthetic test cases to validate our method. We discuss the implications of the synthetic cases and interpret the time-windowed analysis for real data in Section 4. Finally, we conclude about the strengths and limitations of our proposed method in Section 5.
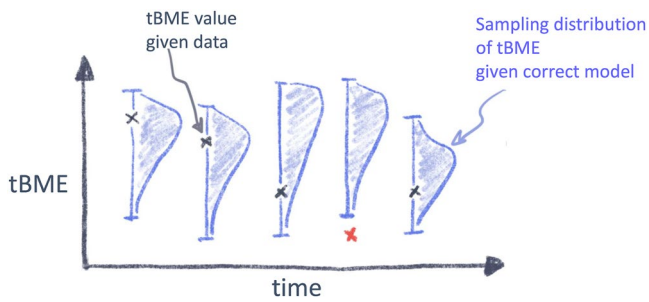
Figure 1. Conceptual illustration of the proposed method: A time series of hypothesis tests with time-windowed Bayesian model evidence (tBME). If the tBME value given data falls within the sampling distribution of tBME, the model can be considered "pseudo-true" during this period (blue crosses); if the value falls outside the sampling distribution, this is an indication of model error (red cross).

2. Methods

To investigate the temporal occurrence of model error, we propose to perform a time-window-based analysis of model performance, with the performance metric being BME. Recall that we understand the model as the set of assumptions about model structure, parameter uncertainty and measurement uncertainty (possibly also input uncertainty). This full definition of a model corresponds to a specific BME value if we evaluate it over the full calibration data time series (traditional Bayesian model evaluation analysis). Now we wish to detect errors that are not covered by measurement noise and parameter uncertainty (i.e., model structural errors, and/or input errors, if input uncertainty is not accounted for by the model setup). Assuming that the model is generally fitting the calibration data well, it should receive a high BME score most of the time; only in periods with missing/misspecified processes acting, the model's performance will deteriorate significantly. This sudden misfit will cause BME to decline, and hence we can use a time-window based BME version of (tBME) as model error indicator.

The challenge is now to distinguish "noisy fluctuations" in BME from statistically significant declines that actually represent model error. Since there is no meaningful interpretation of BME values on an absolute scale, we need a reference to compare the obtained tBME values to. We therefore determine a reference interval that contains hypothetical tBME values under the assumption that the model is in fact true and error-free (the fact that we still see an interval and not just a single value is due to parameter uncertainty in our model). We can then perform a time series of hypothesis testing: If the model is error-free per window, the corresponding tBME value given observed data should fall within the hypothetical interval. This means, at this time scale (of the time window size), the model can be considered "pseudo-true," that is, the model is representative of the true system in a statistical sense.

If the tBME value drops outside the interval, we can conclude that the current model setup does not accurately describe the observed system state in a given time window, that is, that model error occurs. This procedure is schematically illustrated in Figure 1.

In the following, we will explain the theoretical development of our proposed method. Since the method is settled in a Bayesian framework, we start with fundamentals of Bayesian model evaluation in Section 2.1. Then, we explain the moving time-window Bayesian analysis in Section 2.2, mention relevant time scales in Section 2.3, and discuss the theoretical behavior of the tBME curve for varied time scales in Section 2.4. Section 2.5 describes the construction of the reference interval for tBME interpretation. As a side product of time-windowed calibration, we obtain dynamic posterior parameter distributions introduced in Section 2.6.

2.1. Bayesian Model Evidence (BME)

The Bayesian framework allows to model and quantify an uncertain state of knowledge, which can be updated in the light of observed data D. An uncertain quantity of interest y predicted by a model M can be expressed as a random variable with a prior PDF $p(y|M)$, where PDF stands for probability density function. By determining the likelihood of the model prediction to have generated the observed data, $p(D|y, M)$, this prior PDF is updated to a posterior PDF $p(y|D, M)$:

$$p(\mathbf{y}|\mathbf{D}, M) = \frac{p(\mathbf{D}|\mathbf{y}, M)p(\mathbf{y}|M)}{p(\mathbf{D}|M)} \tag{1}$$

The term in the denominator, $p(D|M)$, is a model-specific normalizing constant that marginalizes over the posterior PDF, and represents the model's marginal likelihood. This term describes how likely it is that the model M has generated the observed data D. It is therefore also called Bayesian model evidence (BME), which can be expressed as

$$\text{BME} \equiv p(\mathbf{D}|M)$$

$$= \int_{-\infty}^{\infty} p(\mathbf{D}|\boldsymbol{\omega}, M)p(\boldsymbol{\omega}|M)d\boldsymbol{\omega} \tag{2}$$

$$\approx \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} p(\mathbf{D}|\boldsymbol{\omega}_i, M),$$

where ω contains all uncertain aspects of the model M, and $p(\omega|M)$ is the prior PDF of ω under the model M. The last row shows a brute-force Monte Carlo numerical approximation of BME (Schöniger et al., 2014): A large number N_{MC} of random parameter realizations ω_i is drawn from the parameters' prior PDF $p(\omega|M)$ to form a Monte Carlo ensemble, the corresponding model predictions y_i are determined, and their likelihoods are averaged to approximate the Bayesian integral. This approximation is assumption- and bias-free, but computationally expensive. For other possible approximations of BME, readers are referred to Liu et al. (2016); Oladyshkin and Nowak (2019); Schöniger et al. (2014).

BME is perfectly suited as a model performance metric in the face of uncertainty (in parameters, forcings, and output data), because it balances a model's goodness-of-fit with a penalty for overcomplexity, see Höge et al. (2018) for a review on this topic. To give a simplistic explanation, think of a 1D prediction: BME is exactly the PDF value of the model's predictive distribution read off at the observed data value. But for multiple model outputs, especially for time series, the mathematical dimensionality (i.e., length of the vector y) is much higher. The numerical approximation in Equation 2 avoids the problem of estimating high-dimensional PDFs (as would be necessary for long time-series of calibration data) by sampling the relatively low-dimensional parameter space instead. Thus, it obtains a sample of the predictive PDF by running the forward model on the parameter sample.

As likelihood function to represent measurement error, we choose a Gaussian distribution centered about the observed data D:

$$p(\mathbf{D}|\boldsymbol{\omega}_i, M) = \frac{1}{\sqrt{(2\pi)^{N_o}|\mathbf{R}|}} \exp\left(-\frac{1}{2}(\mathbf{D} - \mathbf{y}_i)^T \mathbf{R}^{-1}(\mathbf{D} - \mathbf{y}_i)\right), \tag{3}$$

with N_o representing the number of observations (i.e., the length of D). In this assumption, the model misfits (residuals) $\mathbf{D} - \mathbf{y}_i$ are attributed to uncorrelated measurement errors, that is, error covariance matrix R of size $N_o \times N_o$ has non-zero elements only on its main diagonal. It is clear that model errors will show a complicated pattern of correlation (see, e.g., Evin et al., 2013), and many authors used the likelihood function to also represent model errors (Ammann et al., 2019; Smith et al., 2015). We are not doing so, because we do not want to prescribe any specific model error pattern a priori. Instead, we let the data speak and identify the model errors for us. We therefore propose to start with the simplified assumption of uncorrelated errors in the model error detection phase, and recommend to use the result of this analysis as input for more realistic modeling of model errors (or even better: improvement of the underlying conceptual or physics-based model) in a subsequent phase of model error mitigation. A modeler could also treat random measurement error and known systematic model error in a lumped fashion, in order to detect further yet unknown errors. Besides, the choice of likelihood function is not limited to the one we show here, and does not influence the functioning of our framework.

2.2. BME With a Moving Time-Window (tBME)

Recall that we assume that the model under investigation is generally correct except for some unresolved processes that occasionally occur. Under this assumption, the goodness-of-fit of a model given a short dataset should be high, unless this subset of the calibration time series includes the occurrence of unresolved processes. Based on this concept, we propose to calibrate the model with a moving window, using BME as a model performance metric.

The analysis starts with the selection of a fixed window size. This window size determines the length of the time-series subset chosen for BME evaluation. With a time-window size of τ, BME is evaluated on a dataset of τ time steps. Equation 1 then rewrites as:

$$p(\mathbf{y}_{T_j(\tau)}|\mathbf{D}_{T_j(\tau)}) = \frac{p(\mathbf{D}_{T_j(\tau)}|\mathbf{y}_{T_j(\tau)})p(\mathbf{y}_{T_j(\tau)})}{p(\mathbf{D}_{T_j(\tau)}|M)} \tag{4}$$
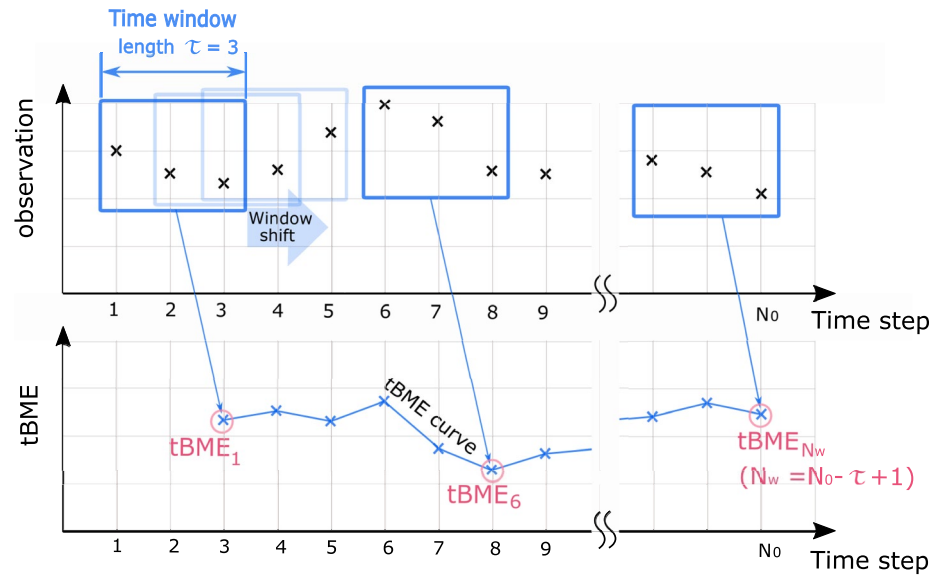
Figure 2. Illustration of time-windowed time-windowed Bayesian model evidence (tBME) evaluation with a moving time-window. A series of BME evaluation constructs the tBME curve, which is later used as an indicator for model structural error occurrence.

which transforms BME (Equation 2) to be tBME and writes as:

$$\text{tBME} = p(\mathbf{D}_{T_j(\tau)}|M). \tag{5}$$

Note that evaluating $p(\mathbf{D}_{T_j(\tau)}|M)$ according to Equation 2 means to apply the likelihood function (Equation 3) to the data subset $\mathbf{D}_{T_j(\tau)}$ instead of D.

This time window is then sliding through the observed time series to form a tBME curve as a function of time, which we call tBME curve (see Figure 2). We obtain a number of N_w individual tBME values that make up this curve. N_w is equal to ra, that is, the first time window "eats up" τ data points of the original data time series, and then this window is shifted time step by time step ($j = \tau, \tau + 1, \tau + 2, \ldots, N_o$) until the final window contains the last data point with index N_o (cf. Figure 2). The choice of shifting by a certain time step (e.g., typically a day in hydrology) yields the maximum number of possible windows; computational time constraints could justify a shifting of more than one time step at once and thus reducing the number of tBME values to be evaluated. The length of the time step can be arbitrary, depending on the frequency of the measurements. In our application (3.4) we choose 1 day for one time step.

2.3. Relevant Time Scales for tBME Analysis

Before discussing the different time scales that are relevant for the tBME analysis, we wish to differentiate between the terms error period and residual period: With error period, we will refer to the time span during which model error occurs (e.g., model lacks a snow-melt routine), while we define the residual period as the time span during which the effect of model error can be observed in form of model residuals (e.g., modeled soil conditions are too dry because input from snow-melt is missing). Therefore, the length and position of residual periods does not necessarily coincide with the length and position of the error period itself; instead, residual periods are typically longer in hydrological modeling because errors tend to accumulate over time, and they might appear with a time delay. Since we do not know the source of model error in real-world settings, we can only detect residual periods with the proposed method and hypothesize about the responsible error periods based on our expert knowledge.

Besides the residual period with length L_e, three additional time scales are relevant in our proposed method: the window size τ, the residual recurrence interval R, and the total dataset length N_o (see Figure 3). The term "residual recurrence interval" stands for the residual-free time interval between two distinguishable residual periods. Note
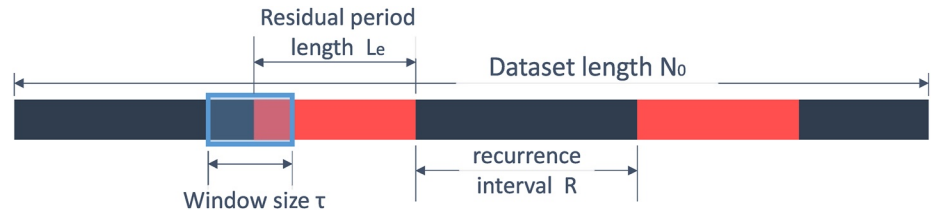
Figure 3. Schematic illustration of the four time scales relevant in time-windowed Bayesian model evidence analysis.

that the two residual periods in Figure 3 do not have to be caused by identical sources. The only condition for errors to be detectable is simply the residuals' significance.

The observed dataset should be long enough to at least contain a complete residual period L_e. Further, the residual recurrence interval R should be larger than the window size τ, so that the residuals caused by separate events/ hydrological conditions are not included in the same time window at any point in time. This ensures identifiability of multiple error segments as individual residual periods. Nonetheless, we will demonstrate the usefulness of our proposed method even for superimposed errors (Section 3.4), since this is a phenomenon often anticipated in real-world applications.

2.4. Theoretical Behavior of tBME Curves

Figure 4 schematically illustrates the theoretically expected behavior of tBME for different window size τ and residual period length L_e. The sketch at the top depicts how tBME curves behave generally with and without parameter compensation. In each of the four panels below, the top row identifies the respective case as a function of the length of the time window τ (blue box) relative to the residual period length L_e (red-shaded box). The plots illustrate the corresponding theoretical behavior of tBME curves. The bottom row summarizes the relationship between the total tBME signal length, denoted as L_s, time-window length τ, and residual-period length L_e. We provide exemplary numbers for L_s, with one unit corresponding to half a gridline increment. In each case, the blue window slides from the residual-free period (high tBME) through the residual period of length L_e into the next residual-free period (high tBME again). We assume here that residuals are pronounced enough to produce a clear signal, and that conditions are exactly the same before and after the residual period, so that the tBME curve will climb back to its previous level.

In each panel of Figure 4, four major states of the theoretical tBME curve can be identified, as the blue time window would slide along the solid red residual period:

1. State I: The time window is still outside of a residual period and produces a high tBME value.
2. State II: The time window starts to contain the residual period and tBME drops to a low value.
3. State III: For the time during which the time window completely captures the residual period, the tBME curve will remain at a low plateau. Potentially, if parameters can successfully compensate model error, tBME temporarily returns back to its previous high value (state III*).
4. State IV: The time window starts to move into a residual-free period again and the tBME curve rises to the pre-residual value (state I).

Error signals are defined in State II, when a tBME curve drop can be recognized. It occurs when the time window partially covers the residual period. At this state, the observed data points inside the window are obtained from two different system conditions. Even if there are realizations that fit segments of the data well, there still exists another part of the dataset in the time window that cannot be fitted with the same parameter values. As a consequence, goodness-of-fit is reduced for that window, and tBME drops.

In the following, we discuss the differences between the four cases distinguished in Figure 4. If we assume that parameters cannot compensate for model error at all (there is no parameter realization in the model ensemble that can fit the data well during the residual period), we see a single drop to the low tBME value, no matter whether the window is longer or shorter than the residual period (following the red lines to the low tBME level in Figure 4). This is clearly visible in case 1, 2, and 4. The relationship between two specific time scales defines how
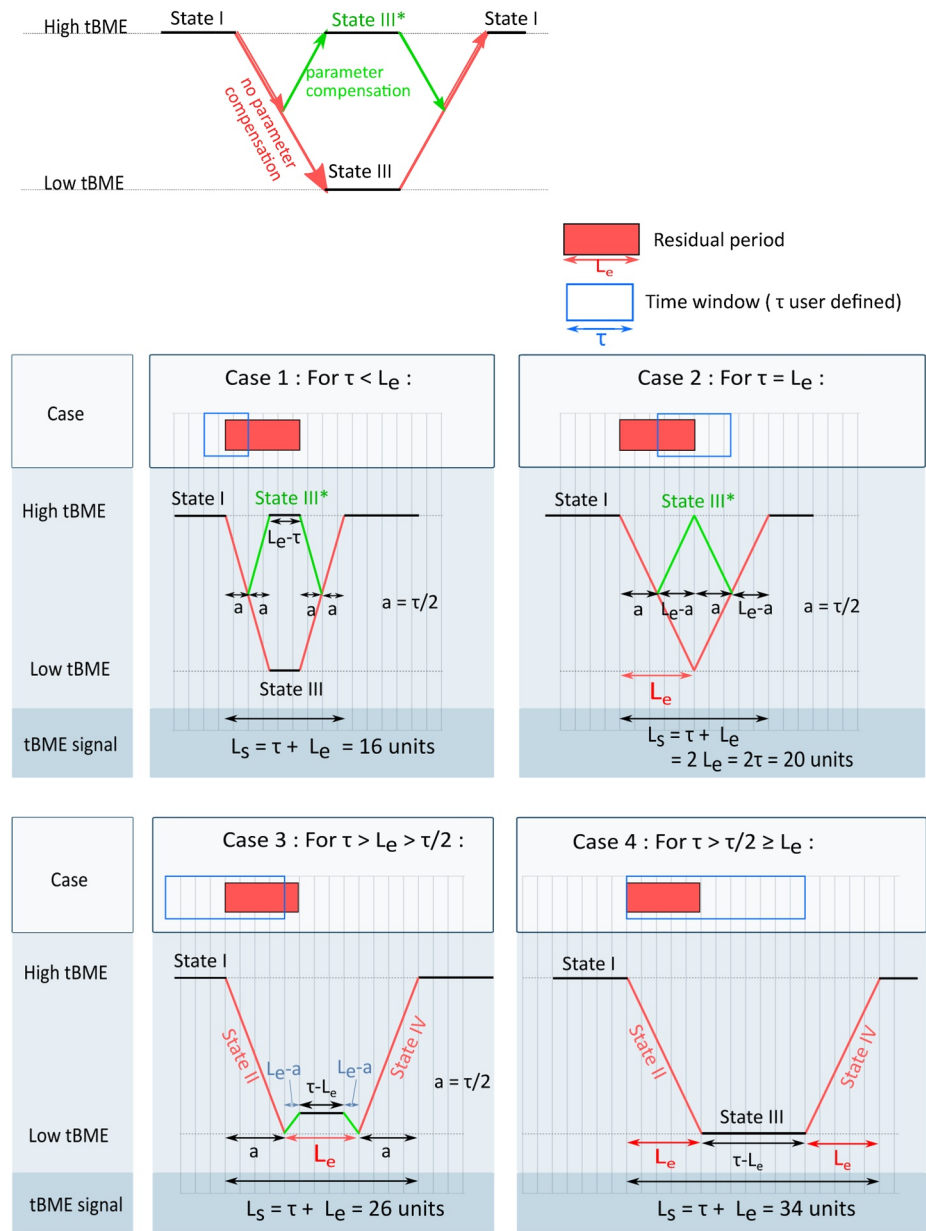
Figure 4. Schematic illustration of expected time-windowed Bayesian model evidence signal as a function of residual period length L_e (here: Fixed at 10 exemplary units) and window size τ (here: Varied from 6, 10, 16, to 24 units).

long the low tBME value will persist: it can be a valley of length $L_e - \tau$ for $\tau < L_e$ (case 1), it can be a single inflection point in the case of $\tau = L_e$ (case 2), or it can be a valley of length $\tau - L_e$ in the case of $\tau > L_e$ (in case 3 and 4).

To some degree, modified parameter values might be able push the model's predictions back into the right direction (for the wrong reasons): parameters have a tendency to "absorb" model error. The degree to which this effect is happening depends on the flexibility of the model structure and the type of error. If the model parameters are able to compensate for model error, tBME may temporarily return to its previous high value (state III*). In this state, the observed data cannot be fitted well by the model with the original parameter set, but the model still can describe the data well with other parameter values. This hypothesis suggests that the tBME curve will recover within a residual period, if the data before and during the residual period may have been generated under the same "model setup." Such type of model error could be treated well with, for example, time-dependent parameters. Generally, this state is somewhat dangerous, because it might be interpreted as the end of the residual period,

although in fact conditions have not changed. It is therefore important to keep an eye on the further development: Our basic assumption is that, during most of the time, there is no significant model error, so typically a stable residual-free period should follow after a tBME climb. If, instead, we observe a second dive right after, we should be warned and suspect state III* instead of state IV.

The timing of the temporary tBME rise in state III* can be determined as follows: when the residual period covers half of the time window, this is the worst state in terms of model performance - because parameters have to fit two conflicting states (error-free and error-prone model) simultaneously, which is impossible to achieve. As soon as the window slides further into the residual period, error conditions take up a larger fraction in tBME calculation, and now parameter compensation can start to come into effect, so the tBME value has a chance to increase. It may rise until the window covers the full residual period, and it will stay at this high level until the window starts leaving the residual period. This results in the fact that, just like the lower valley, also the upper intermediate peak (state III*) might turn into a plateau of different length, depending on the relationship between window size τ and residual period length L_e (see Figure 4).

Finally, once the window has moved out of the residual period completely, the tBME curve fully recovers, that is, it returns to its pre-residual-period high value and remains there in a stable fashion (state I).

We find that the total signal length L_s equals $L_e + \tau$, no matter if the window size τ is longer or shorter than the residual period length L_e. Hence, by observing the total signal length L_s from the analysis with a specific τ, we can directly infer the residual period length to be $L_s - \tau$. This has a large practical value and will help our interpretation of tBME curves over varied window sizes. We will illustrate and discuss this further on our case studies in Section 3.

What we did not take into account when drawing this schematic figure is the fact that tBME calculation will yield larger differences between high and low values for larger datasets, that is, for larger τ. Therefore, the signals become more evident from larger τ. Finding an optimal window size τ is therefore a compromise between a long enough window to obtain a clear signal, a short enough window to only capture individual error events and avoid "smearing" of different types of model error, and a short enough window to enable fast convergence of the tBME sampling distribution.

In practice, we find that measurement noise and other overlaying effects may distort a clear error indication for window sizes smaller than the residual period length. Our investigations suggest that increasing the window size gradually enlarges true error signals and is hence advisable for reliable error detection (Section 4.1). This is apparent from Figure 4 especially for $\tau/2 \geq L_e$ (case 4): The signal is enlarged to a wider valley, because effective error compensation by parameters is not possible anymore, since the residual-free period always takes up a larger fraction within the moving window than the residual period. Shifting parameter values would lead to a lower performance in reproducing the error-free data, and hence the overall skill metric (here: tBME) would decrease. Naturally, there is an upper limit to increasing the time-window size: once we arrive at $\tau = L_d$, there will be only a single window left, tBME will be equal to BME, and we will not have any error signal. When moving from $\tau \approx L_e$ toward this upper limit, the signal will start to become weaker at some point because the residuals within the residual period are diluted too much in the residual-free data. Additionally, the computational effort becomes increasingly challenging. Hence, we recommend to test window sizes starting from the anticipated temporal scale of noise and going up to roughly $\tau \approx 2L_e$.

2.5. Sampling Distribution of tBME as Reference

To distinguish a "low tBME value" from a "high tBME value" as simplistically labeled in Figure 4, we wish to construct a reference for the obtained tBME values. If a tBME value falls into the model's "personal BME range," the model's behavior in this time window j is interpreted to be quasi-true; if the tBME value falls outside of that range, statistically-significant model error is detected.

To obtain this range, we randomly sample tBME per time window by randomly picking a model realization as modeled observation data $\mathbf{D}_{T_{j(\tau)}}$ from the model ensemble. We then evaluate the corresponding tBME value by brute-force integration with the complete ensemble minus the picked realization, that is, with an ensemble of size $N_{MC} - 1$. Repeating this leave-one-out procedure with the full ensemble (and hence representing the full predictive uncertainty in the model) will yield a tBME distribution that reflects the predictive variability in the

model. This variability is typically due to measurement noise and parameter uncertainty (which is the source of uncertainty that we consider in our case study in Section 3), but other sources of uncertainty could as well be accounted for by the tBME reference distribution. For example, known input uncertainties could be accounted for when constructing the output ensemble, and hence the tBME reference distribution would reflect this additional source of uncertainty by sampling the model output as described above.

With the tBME sampling distribution as reference, we perform a series of hypothesis tests: for any specified position of the moving window, does the tBME value given real data fall within the high-density region of the tBME sampling distribution, or is it at the low-BME low-probability tail, or does it even fall outside of the sampled range? It is in the modelers hands to decide for a significance threshold to reject the hypothesis that the model has generated the data (e.g., tBME lower than the xy^{th} quantile of the sampling distribution, or tBME smaller than the smallest sampled BME, etc.). We do not give a general recommendation here, because this significance threshold should be chosen context- and study-specific. We will discuss this further during the interpretation of our test case results, see Section 4.

By observing the dynamic trend of the tBME curve compared with the tBME reference distribution, we can monitor how model performance varies and thus diagnose temporal model error occurrence as shown in Figure 1. Note that the tBME curve and the reference distribution should be evaluated with the time window of same length to ensure a consistent comparison. In this study, reference tBME distributions remains the same for all test cases, because the model setup (including all assumed distribution) is unchanged.

2.6. Dynamic Posterior Parameter Distribution

When a model is calibrated with a sliding time-window, a series of posterior parameter PDF samples are obtained "on-the-fly." If varying conditions in different time windows favor significantly different parameter values, we expect to see significant shifts in the posterior parameter samples, for example, visualized as histograms. Such shifts provide information about potential parameter compensation effects.

We recommend inspecting the series of posterior parameter distributions as a diagnostic tool to learn about the type of model error that was detected with the tBME analysis. The dynamics of the posterior parameter distributions may point the modeler the missing processes, wrong assumptions about mass balance components, or similar misspecifications.

2.7. Summary of Proposed Workflow

The following summarizes the steps of the workflow (see Figure 5):

1. Determine τ: Determine the time scale of the error duration that is expected to be investigated.
2. tBME reference distribution preparation.
    2.1. Draw a random realization from the model ensemble as modeled observation data time series, and evaluate tBME values with the time window.
    2.2. Repeat step 2 with another realization.
    2.3. Determine the tBME confidence interval.
    2.4. Check convergence of tBME confidence interval.
    2.5. Repeat from step 2.1 until convergence is fulfilled.
3. tBME analysis: Evaluate tBME curve with observed data.
    3.1. Evaluate tBME curve given τ and real observation data.
    3.2. Plot tBME reference distribution and tBME curve given real observation data.
4. Repeat from step 1 with variation of the time window size τ.

3. Demonstration on a Hydrological Case Study

We demonstrate our proposed approach to model error detection on a hydrological case study from the field of soil hydrological modeling. We first introduce the model and field data in Sections 3.1 and 3.2, then explain the numerical implementation in Section 3.3, define synthetic test cases for our analysis in Section 3.4, and finally present and discuss the results in Section 4.
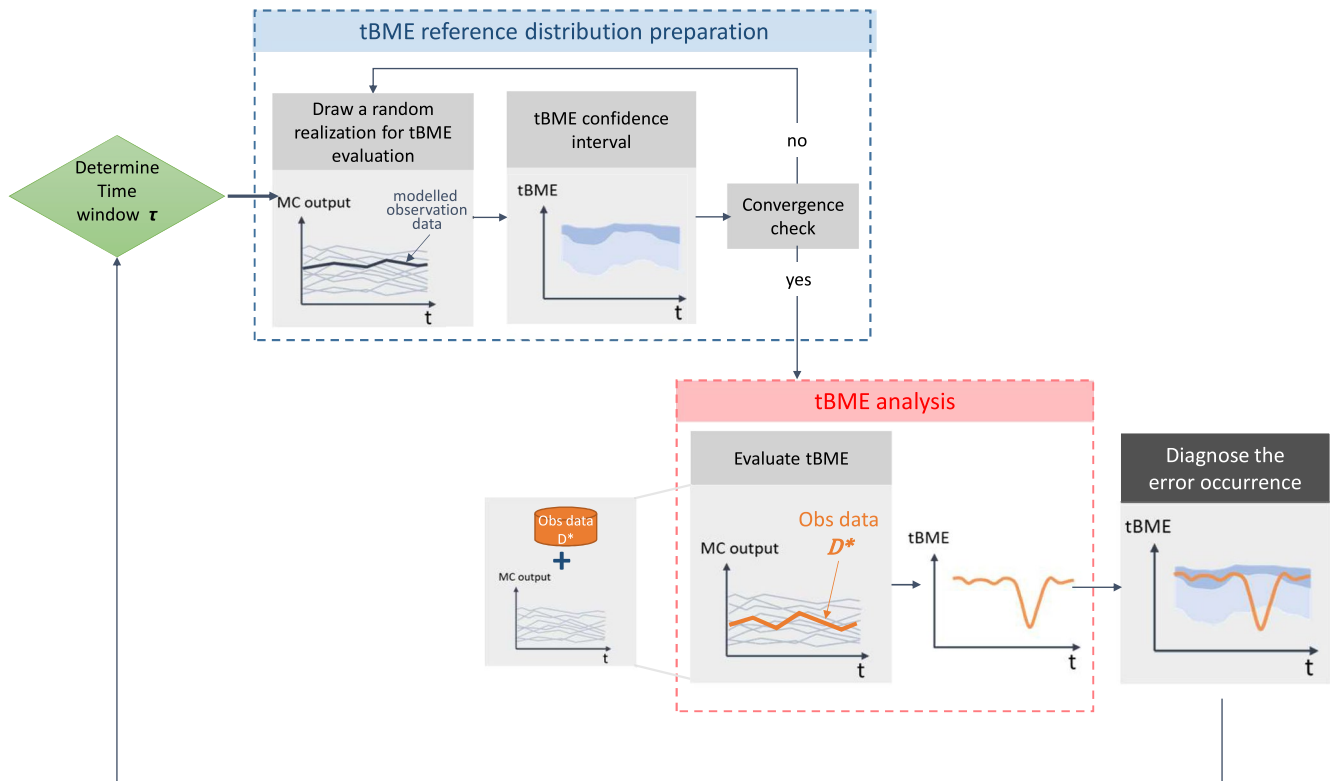
Figure 5. Diagram of the proposed workflow of tBME method analysis. In the part of tBME reference distribution preparation, a randomly drawn realization is used as data D (we call this D "modeled observation data"). In the error diagnosis (tBME analysis), the real observation data D* is used for the analysis.

3.1. Soil Hydraulic Model

We use HYDRUS-1D (Simunek et al., 2005) to numerically solve the Richards equation, which describes water movement in a one-dimensional vadose zone. It writes as:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z}\left[K\left(\frac{\partial h}{\partial z} - 1\right)\right] - R,$$ (6)

where θ is the volumetric water content (−), h is the soil water pressure head (cm), t is time (day), z is the depth (m), K is the unsaturated hydraulic conductivity function (cm day^{-1}), and R (cm^{3}cm^{-3} day^{-1}) is a sink term when root water uptake is considered. This equation has three unknowns, namely h, K, and θ, and thus requires two constitutive relations for closure: the water retention function $\theta(h)$ and the unsaturated conductivity function $K(\theta)$. A commonly used description of these hydraulic properties is the Mualem-van Genuchten (MVG) model (Mualem, 1976; van Genuchten, 1980):

$$\theta(h) = \begin{cases} \theta_r + \dfrac{\theta_s - \theta_r}{(1 + |\alpha h|^n)^m} & h < h_s \\ \theta_s & h > h_s \end{cases}$$ (7)

$$K = K_{sat}\Theta^l\left[1 - \left(1 - \Theta^{\frac{1}{m}}\right)^m\right]^2,$$ (8)

with the effective saturation

$$\Theta = \frac{\theta - \theta_r}{\theta_s - \theta_r},$$ (9)

Table 1
Bounds of the Mualem-van Genuchten Model Parameters Used for Uniform Prior Sampling

	θ_s (m³ m⁻³)	α (m⁻¹)	n (−)	K_{sat} (ms⁻¹)	l (−)
Lower bound	0.3	1	1.1	10^{-7}	0.1
Upper bound	0.7	20	9.0	10^{-3}	1

where θ_r and θ_s refer to the residual and saturated water content (−) respectively; α (cm⁻¹) and n (−) are shape parameters, K_{sat} refers to the saturated hydraulic conductivity (cm · day⁻¹), $m = 1 - \frac{1}{n}$ is an empirical coefficient, and l (−) is the pore connectivity parameter by Mualem (1976). We use a modified form of the MVG model with an air-entry value of $h_s = -2$ cm (Vogel & Cislerova, 1988).

In one of the considered test cases (definition see Section 3.4), we use alternatively the dual porosity model by Durner (1994) that describes the retention properties as a linear combinations of the van Genuchten curves for different pore spaces:

$$\Theta = \sum_{i=1}^{k} w_i \left[\frac{1}{1 + (\alpha_i |h|)^{n_i}} \right]^{m_i}, \tag{10}$$

where w_i are weighting factors for the subclass of the soil with $w_i \in (0, 1)$ and $\sum w_i = 1$. One of the advantages of the dual porosity model is the similar interpretation of the parameters as in the van Genuchten model (Durner, 1994).

3.2. Study Area and Field Data

The field data used in this research is from the Spydia experiment at site in the northern Lake Taupo catchment, New Zealand. Tensiometric pressure head at five depths was measured every 15 min with tensiometer probes (UMS T4e, Germany, accuracy 0.5 kPa) at five different depths between 0.4 and 5.1 m (Wöhling & Vrugt, 2008), but in this research we only use the field data of the three depths: 0.4, 1.0, 2.6 m. The daily measurements of the top layer (0.4 m, the closest layer to the surface among the three layers) show the most interesting dynamics and are therefore used in this research. Meteorological data from the 500 m-distant Waihora station is applied for potential evaporation with the Penman-Monteith equation (Monteith, 1981), which then used as a boundary condition in Hydrus for the computation of soil evaporation and root water uptake. For a detailed description of the model setup, please refer to Wöhling et al. (2008) and Barkle et al. (2011). Likewise, detailed information about the field experiment is given in Barkle et al. (2011); Wöhling et al. (2009); Wöhling and Vrugt (2008) and is therefore not repeated here. Wöhling et al. (2008) investigated seven different soil hydraulic models and found that none of the models fitted to the field data for all time periods, that is, that all models showed structural error for at least some time period. The study focused on improved predictive performance by Bayesian model averaging or Bayesian model combination (BMA/C) techniques rather than on the identification of model errors. In the spirit of model diagnostics and process understanding, it is therefore interesting to investigate the model deficiencies in more detail.

3.3. Numerical Implementation

In our simulation, we use the model ensemble by Wöhling and Vrugt (2008), which consists of $N_{MC} = 900,000$ realizations to represent uncertainty in the 5 MVG parameters θ_s, K_{sat}, α, n, and l per soil layer, adding up to 15 uncertain parameters for 3 soil layers. θ_r is fixed to 0, because it is not sensitive in the considered calibration setup (Wöhling et al., 2008). The prior ranges for the uncertain parameters are given in Table 1, with all three soil layers sharing the same parameter bounds. As stated in Section 3.2, we focus on the simulations of pressure head in the top soil layer for model evaluation and model error detection.

For a given dataset $\boldsymbol{D}_{T_j(\tau)}$ of a specific time-window size τ, we determine tBME (Equation 5) by brute-force Monte Carlo integration (Equation 2). For this step, we need to define the likelihood function $p(\boldsymbol{D}_{T_j(\tau)} | \boldsymbol{\omega}_i, M)$ (Equation 3). We here assume a measurement error standard deviation of 0.09 cm for the soil water pressure head data. This evaluation is exactly like a brute-force Monte Carlo implementation for traditional Bayesian model evaluation, only that the calibration dataset is deliberately chosen to be very short.

To assess the convergence of our numerical tBME scheme, we determine the effective sample size (ESS) (Liu, 2004). The ESS counts how many parameter realizations significantly contribute to the tBME estimate. We also perform a visual analysis of the tBME convergence behavior. Both analyses together provide a good indicator

of how stable the obtained numerical result is. Although we have monitored tBME convergence in all investigated cases, we wish to emphasize that, for real-world applications, a (too) low ESS is not annoying for two reasons: (a) Low ESS arise if only very few realizations show a non-zero likelihood, that is, if most of the ensemble is performing really poorly. This is the situation that we are in fact most interested in: It reveals severe model error. In this case, we will obtain a low tBME value; and we do not care whether it is low, very low, or extremely low—the numerical tBME estimate does not have to be that accurate. (b) We do not use the obtained tBME value for any further procedure (such as model-averaged prediction), but we are satisfied with tBME serving as more of a qualitative model error detector. We are therefore not relying on an accurate estimate of tBME for further processing.

To now obtain a complete tBME curve for a chosen time-window size τ, we start by evaluating the first tBME value (tBME$_j$ with $j = \tau$) for the period of simulation at days 1 to τ. Then we slide this window by 1 day, determine tBME$_j$ for $j = \tau + 1$, and so on, until we hit the end of the full calibration dataset (cf. Section 2.2). In this way, we construct a tBME curve for a given dataset, for example, for the observed time-series of pressure heads.

We repeat this routine for a number of 30,000 randomly selected output realizations used as modeled observation data, in order to construct the tBME sampling distribution as reference (cf. Section 2.5). Convergence of the sampling distribution was checked visually and by monitoring ESS. Note that ESS does become relevant in this part of the procedure, because we want the reference distribution to be informative and reliable. We therefore ensure that, during these runs, ESS does not drop below the value of 200. This is usually not a limiting factor, because the analysis with the modeled observation data by definition assures that the model is true, so BME performance does not drop that dramatically, and neither does ESS.

Finally, we evaluate the entire time series of tBME values given the observation data with the sliding time-window method. In our study, we analyze the data with four distinct increasing values of window size, that is, $\tau = 5$, 10, 15, 20 days, to create four tBME curves accordingly.

For each time window that we use for calibration (and to determine tBME), we obtain a posterior parameter sample as a "side product." We simply weight our prior parameter ensemble with the corresponding vector of likelihoods (dimension $1 \times N_{MC}$) and approximate the posterior PDF by kernel density estimation on the weighted sample. As an analog to the tBME curve, we obtain a series of posterior parameter distributions that we can inspect for further investigation of the potential source of model error.

All our analyses are implemented and performed in MATLAB$^{\copyright}$ R2019a on a standard desktop computer.

3.4. Synthetic Test Cases

Before applying our proposed method to the real dataset, we wish to validate the method under fully-controlled conditions. We use exactly the same model setup as in the real case (see Sections 3.1 and 3.2), but construct the dataset of "observations" to be used for the tBME analysis manually. To test different types of model error that could occur in reality, we define three synthetic scenarios where we knowingly introduce model error. The goal is to confirm that our method will successfully identify these predefined error periods; and thereby we wish to demonstrate how to interpret the model error indicator. Then, we finally use real data to test our approach under realistic conditions.

To generate the synthetic datasets for our analysis, we extract a single realization of model output (soil water pressure head) from the ensemble (base case with no model error, see definition below). In the next step, we manually perturb this realization with different types of model error to generate synthetic datasets that deviate from the model behavior and, hence, model error shall be detected by our tBME method. The perturbations aim to mimic scenarios when a model fails to resolve a temporarily occurring process in the natural system.

While, philosophically, model error is of course in the model predictions and not in the observed dataset, we here pragmatically introduce "errors" to generate the synthetic datasets. This approach is more straight-forward to implement, easier to interpret, and allows us to modify and arbitrarily increase the complexity of the introduced errors without having to set up a new model ensemble for each test case. Whether we introduce the error in the dataset or in the ensemble of predictions does not have an impact on the functioning of our proposed approach, because both ways lead to the exact same residuals.
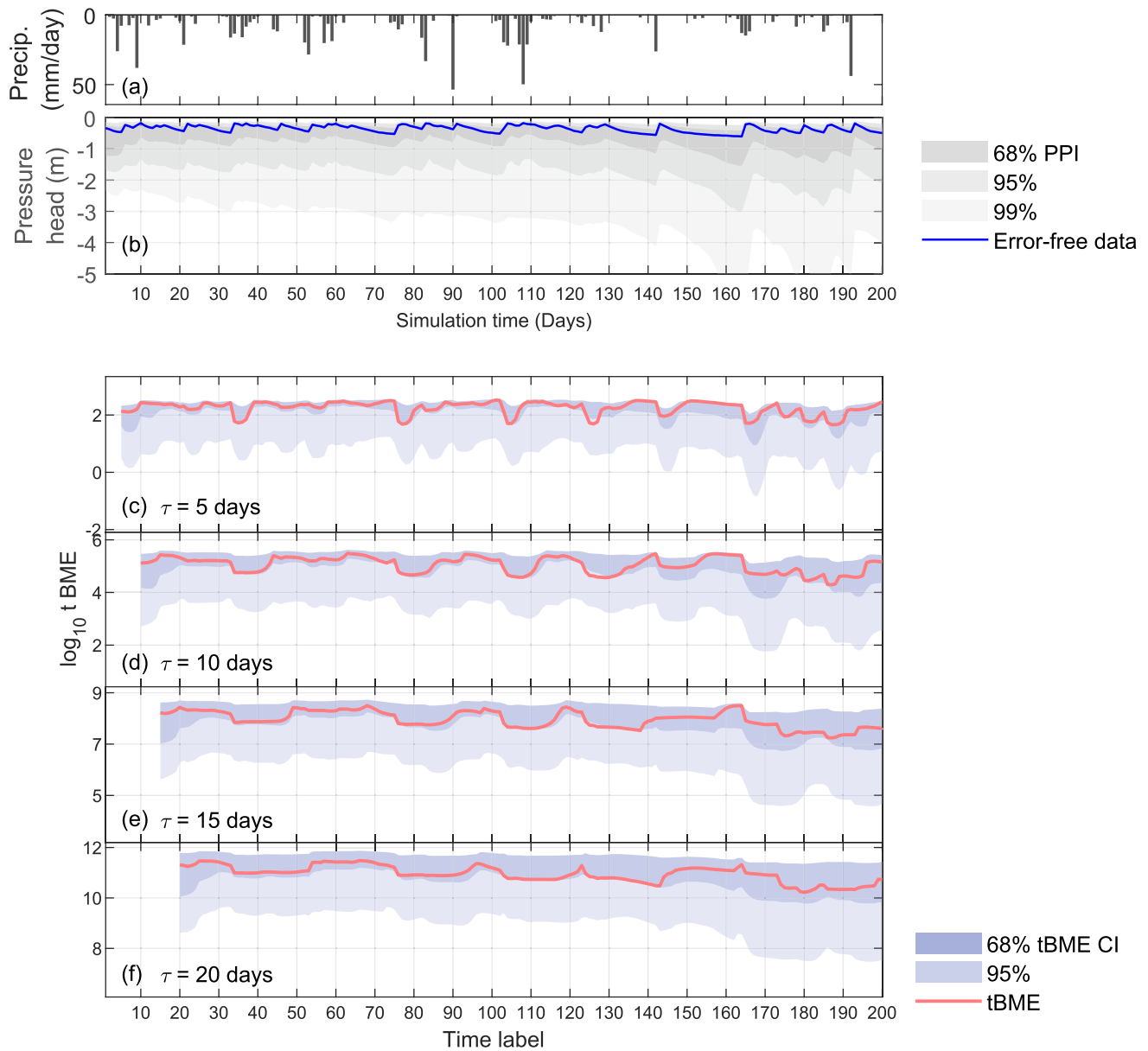
Figure 6. Results for tBME analysis if the model is correct (error-free). 68% PPI denotes prior prediction interval; 68/95% tBME CI denotes 68/95% confidence interval.

3.4.1. Base Case: No Model Error

To demonstrate how our method reacts to the ideal case of no model error, we pick one realization of the model ensemble as the observation data for the tBME analysis. This realization lies within the high-density range of the predictive prior to represent "typical model behavior" (Error-free data in Figure 6b). The model parameters of this case are listed in Table 2. The layers 1–3 refer to the depth of 0.4, 1.0, 2.6 m, respectively. This dataset is considered as a reference case for comparison for the following synthetic cases 1 and synthetic case 2. For the synthetic case 3, we pick a realization which lies within the low-density range of the predictive prior to represent a

Table 2

Mualem-van Genuchten Parameter Values Chosen to Represent Error-Free Data (Base case)

	θ_s	α	n	K_{sat}	l
Layer 1	0.5455	12.9879	4.1150	0.0009	0.1945
Layer 2	0.4783	13.3246	3.6499	0.0006	0.5233
Layer 3	0.5278	5.2775	2.8336	0.0001	0.8315

Table 3
Mualem-van Genuchten Parameter Values Applied to Perturb the Base-case Dataset During Error Periods in the case of Superimposed Errors (case 3)

	θ_s	α	n	K_{sat}	l
Layer 1	0.4146	6.1609	2.5970	0.0001	0.4049
Layer 2	0.3860	2.2569	5.5489	0.0004	0.3108
Layer 3	0.5719	2.9203	4.5101	0.0008	0.9006

"low-probability synthetic truth." Also, this parameter set (Table 3) produces pressure head data with similar dynamics to the field data.

3.4.2. Synthetic Case 1: Error in Model Structure

The observation dataset (Error-free data in Figure 6b) of the base case is now perturbed manually by structural error during specific time periods. We assume that "reality" behaves like a dual-porosity model during these periods (with strong precipitation events kick-starting this behavior, and approximately like a single-porosity model otherwise. Technically, we run HYDRUS-1D with the error-free setup described for the base case, and switch to the dual-porosity model setup during 3 selected error periods: Days 31–40, 81–90, and 161–170 (cf. Figure 7a). The synthetic dataset that is used for tBME analysis is depicted with the label "perturbed data" in Figure 7b.

The following points are considered for selecting these error periods: (a) The error periods do not start before $t = 20$, which is the largest window size used in our method demonstration. (b) The error period length should
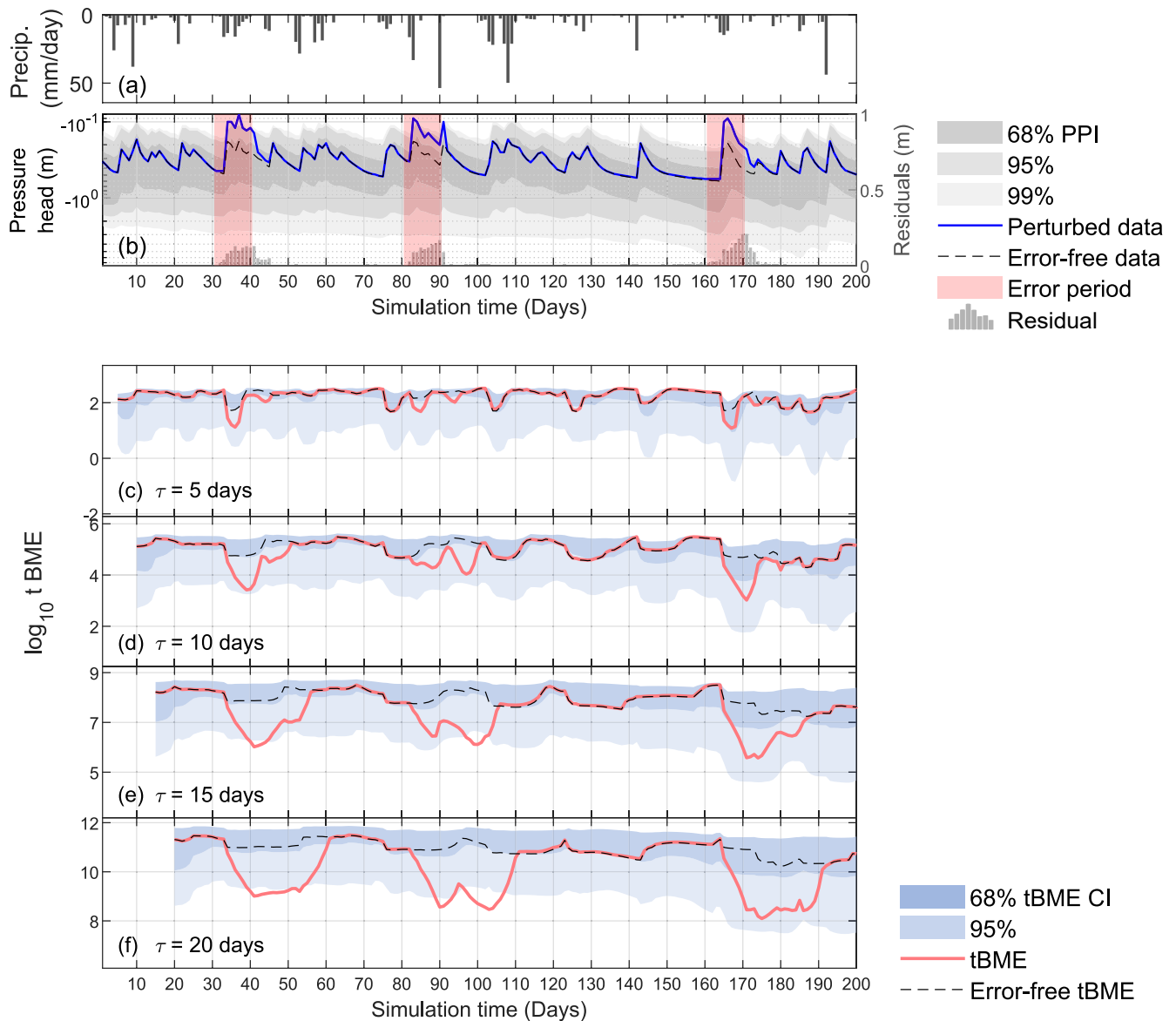

Figure 7. tBME curves given model structural error (synthetic case 1: Reality switches to dual-porosity behavior, model stays in single-porosity setup).

be between the selected smallest and the largest window sizes. (c) The error periods cover simulation periods of different spread in the predictive prior. (d) The largest resulting residual per day (i.e., each bar in Figure 11b) is not larger than 0.2 m.

For constructing the perturbed dataset, the parameters K_{sat}, θ_s and l remain the same as in the base case; parameter values for the first pore subsystem are chosen as in the base case, and parameter values for the second pore subsystem are given in Table 4. Here, the larger values of α and n than in the base case imply that a macropore system is considered as the second soil subsystem. Note that we assign the same dual-porosity parameter values to layers 1 and 2, while layer 3 is assumed to not show a dual-porosity behavior at all.

The ensemble that we test remains the same throughout all test cases: it consists of single-porosity runs (MVG model) with prior parameter ranges given in Section 3.3. Hence, we see significant pressure head residuals during error periods and beyond (Figure 7b), because the tensiometric pressure responses in a dual-porosity soil to heavy precipitation in "reality" (dual-porosity model) is quite different from what our tested single-porosity model predicts.

3.4.3. Synthetic Case 2: Error in Model Forcing

As a second scenario, we consider error in the model forcing. We mimic local effects that are quite common in reality: The nearby meteorological station has recorded precipitation, but the event is so strongly localized (e.g., convective shower), that it actually did not rain at the experimental site. As a consequence, soil moisture measurements show no response (or rather, pressure head becomes even more negative due to the drying soil, see Figure 11), while our model predicts increased soil moisture due to a heavy precipitation event (pressure head becomes less negative, see Figure 11). Technically, the synthetic dataset is generated by running HYDRUS-1D in the setup of the error-free base case (Table 2), but the input of precipitation is removed within the error periods (days 36–40, 89–90, and 166–170). This causes significant residuals during error periods and beyond due to the "storage memory" of the soil hydraulic model, see Figure 11b. Note that, here, the time of the error period (forcing error of, e.g., just 1 day) is much shorter than the residual period (time span during which the impact on simulations is obvious from the residuals, approximately 10 days).

3.4.4. Synthetic Case 3: Superimposed Errors in Model Structure and Forcing

With the third test case, we aim to study the impact of realistic, superimposed errors. We run HYDRUS-1D in single-porosity mode with the parameter set in Table 3. Recall that we pick this realization that lies within the low-density range of the predictive prior to represent a "low-probability synthetic truth" for this synthetic case. During the error periods defined in case 1, we apply the dual-porosity model by switching on w_2, α_2, and n_2 as shown in Table 5. Forcing errors are then superimposed by removing precipitation input on simulation days 44–45, 81, 82, 84, 161, 163, 164, and 165. These superimposed error periods also cause superimposed residual periods.

This leads to a combined impact on residuals between the synthetic dataset and the model ensemble predictions on different time scales, see Figure 15b. Both residuals tend to partially cancel out, such that net residuals are temporarily reduced: Switching to dual-porosity mode means considering macro-pore flow, which compensates the missing precipitation since water moves quicker to greater depths just like it does when it rains.

4. Results and Discussion

We first present the results of the tBME analysis applied to the error-free base case, then use our findings for the interpretation of the results of synthetic test cases 1–3 (Section 4.1), and finally show the results of applying our proposed approach to the observed field data in Section 4.2.

4.1. Results of tBME Analysis for Synthetic Test Cases

4.1.1. Base Case: tBME Given a Correct Model

The base case illustrates what the tBME curve should look like when the model is correct. Figure 6a records the precipitation (model input), and Figure 6b presents the model's prior predictive distribution (with the gray-shaded areas indicating the 68%, 95%, and 99% prediction intervals) and the error-free dataset (blue) used for evaluating

of tBME values. Figures 6c–6f show the tBME curves (red curves) resulting for four different time-window sizes τ. Their corresponding tBME sampling distributions are shown as blue-shaded areas indicating the 68% and 95% confidence intervals. Note that we label these as "confidence intervals," because the resampling procedure invokes a frequentist perspective, although our analysis is otherwise embedded in a Bayesian framework. The x-axis of the tBME curve and the tBME sampling distribution is shifted by τ days. For example, the first tBME value of Figure 6c is plotted at $t = \tau = 5$, which is evaluated with the observed data from the first day to the fifth day (cf. Figure 2).

4.1.1.1. Model-Internal Variability of tBME

As shown by the red curves in Figures 6c–6f, all tBME curves fall inside the range of their tBME sampling distributions during all time periods, which confirms that no significant model error occurs. This meets our expectation, since this dataset is generated by the model. But the necessity of working with time-dependent confidence intervals of tBME as reference becomes apparent. tBME has its own dynamic, and without reference tBME would not be interpretable.

For example, in Figure 6c, the tBME curve for $\tau = 5$ contains various drops. This is due to the model's parameter uncertainty: Even though the base-case dataset is a model realization sampled from the high-density range of the predictive prior, other realizations in the ensemble do not necessarily follow the exact dynamic trend of this individual realization. This effect can be observed when the model is calibrated with a small window size. With increasing time-window size, those small deviations become less influential, because the bias from short-term effects is diluted over a larger window size. This can be observed in the tBME curve for $\tau = 20$, which is smoother than the one for $\tau = 5$. The important conclusion is: no matter how wiggly the tBME curve, if it stays within the tBME confidence range, there is no substantial (statistical) reason to suspect model error - it can just be due to parameter (or other sources of) uncertainty.

4.1.1.2. Dynamics of the tBME Sampling Distribution

It is worth noting that the effects of parameter uncertainty are increasing over time in our case study (Figure 6a), leading to wider tBME sampling distributions at later times (Figures 6c–6f). This demonstrates that the criterion to trigger an error signal (falling below a certain quantile of the sampling distribution) is dynamic in time.

4.1.1.3. tBME Threshold for Error Detection

The challenge is hence to identify a meaningful lower threshold for tBME to label a drop as "model error occurrence." As noted in Section 2.5, there cannot be a general guidance for all cases, but rather we recommend to perform this base-case test before interpreting the results of real tBME curves as done here: by choosing a realization from the high-density region of the predictive prior with a small window size, modelers can check to which quantile of the sampling distribution the tBME curve drops. This level is the highest one that should be used for model error detection (to avoid false-positive error detection); rather, we would expect to see much more pronounced drops with real data if severe model error occurs. Looking at increasing window sizes τ will help gaining confidence in the choice of the threshold. This is discussed in more detail in the context of test case 1 in the following Section.

4.1.2. Case 1: tBME Given Model-Structural Error

Our first synthetic test case demonstrates the time-windowed BME analysis to detect model-structural errors. Figure 7 presents the results in the same order as for the base case (Figure 6).

The red-shaded boxes in Figure 7b represent the error periods during which the "truth" is simulated with the dual-instead of the single-porosity model (see Section 3.4). These time periods now feature an "observed" process that is not contained in the model ensemble.

The corresponding synthetic dataset is shown by the solid blue line (perturbed data in Figure 7b). The base-case dataset is shown as dotted line for comparison. The difference between these two datasets (i.e., the residuals between the model realization corresponding to the base case dataset and the dataset now used for case 1) is plotted as Gy bars. These residuals provide an impression of how wrong the model ensemble is compared to the given dataset (with variations among model realizations due to parameter uncertainty). Recall that we use the term residual period to clearly distinguish this time span of error-induced residuals from the actual error period.
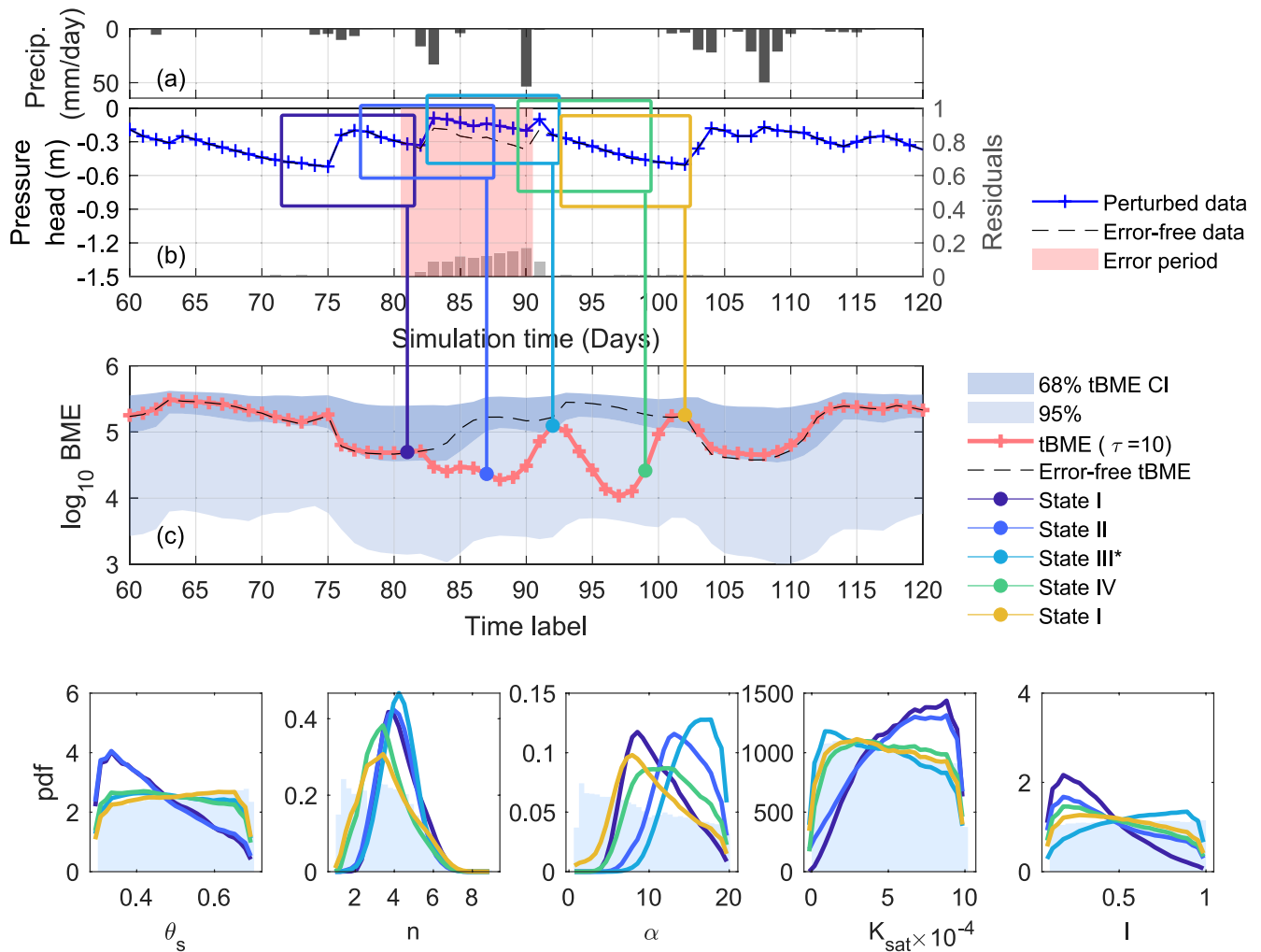
Figure 8. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 10$ (synthetic case 1).

It becomes apparent that the first two error periods only really have an effect once precipitation occurs at $t = 32$ days and $t = 82$ days. During the third error period, residuals are accumulating and are intensified during moderate precipitation from $t = 165$ days. Note that the residuals persist longer than the actual time span of error (residual period longer than error period) due to memory effects caused by water storage.

4.1.2.1. Error Detection

Figures 7c–7f present the tBME curves for the four chosen window sizes. The tBME curves for $\tau = 5$ and $\tau = 10$ show various small drops, which are still within the bounds of the corresponding tBME sampling distributions. This means, the model calibrated on periods of five or ten days still fulfills the statistical condition of the model to be quasi-true. With increasing window size, the signal of error occurrence becomes more pronounced. At t around 40 and 100 days, the tBME curves of $\tau = 15$ and $\tau = 20$ both leave the high-density interval and even the complete range of the tBME sampling distribution, which indicates significant error occurrence. Ideally, a clear turning point can be observed when the time window enters an residual period, as shown by the immediate decline at $t = 33$ and $t = 83$ days in the tBME curve for $\tau = 20$ (Figure 7f).

By inspecting the tBME curve in more detail (Figure 8, for the example of $\tau = 10$), we identify the four major states of the tBME curve introduced in Section 2.4:

1. State I: The time window is outside of a residual period and the tBME curve stays within the high-density interval of the tBME sampling distribution.
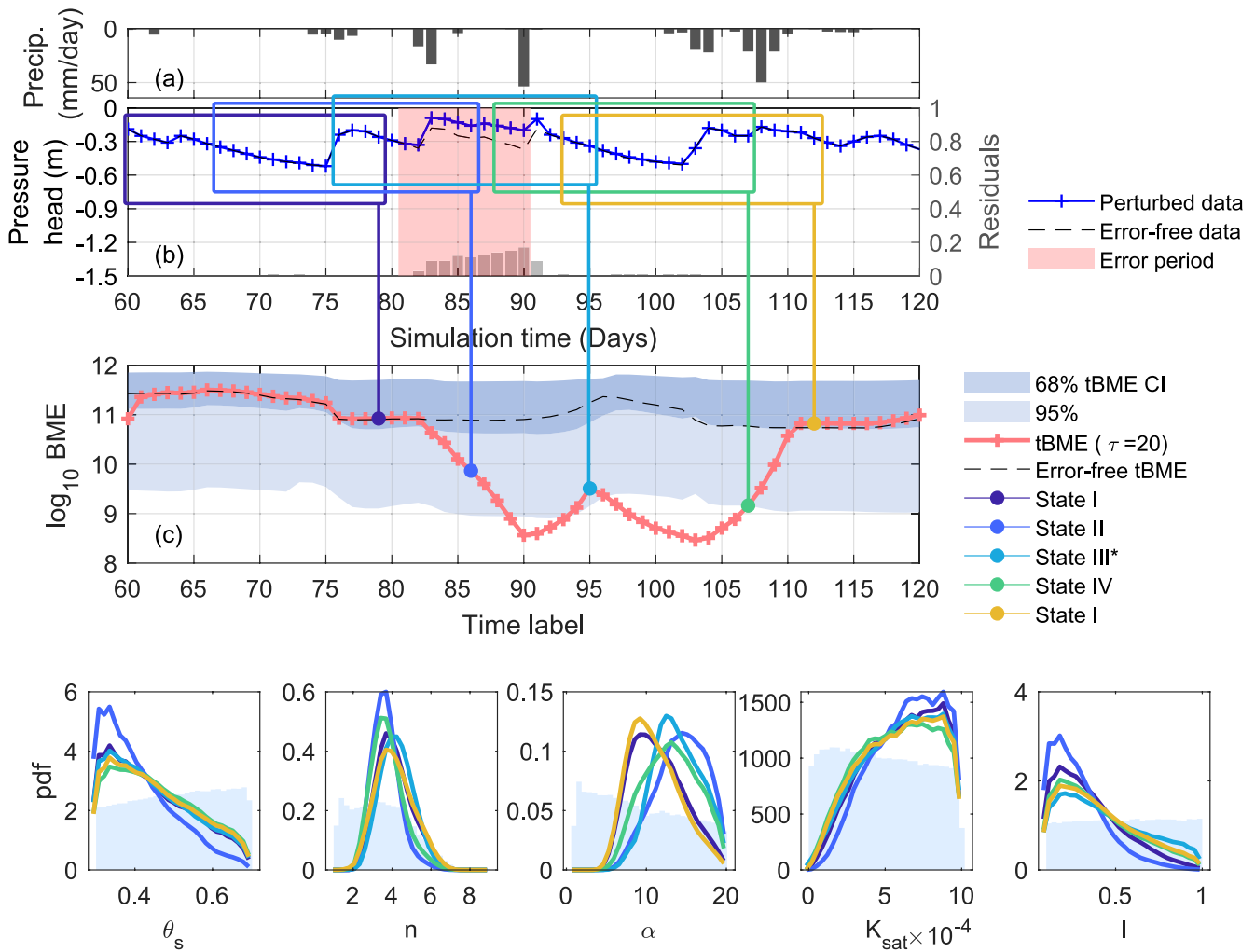
Figure 9. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 20$ (synthetic case 1).

2. State II: The time window starts to contain the residual period and the tBME curve declines.

3. State III*: For the time during which the time window completely captures the residual period, tBME temporarily returns back to the high-density interval of the sampling distribution. This shows that the parameters successfully compensate model error.

4. State IV: The time window starts to move out of the residual period (residual-free) and the tBME curve rises to the high-density interval of the sampling distribution.

For easier recognition, we have marked and color-coded the respective time windows and corresponding tBME values in Figures 8b and 8c. The same four states can be identified for window sizes larger than the perceived residual period length (e.g., for $\tau = 20$, Figure 9), with the only difference that state III* (compensation) is less pronounced. This is because the time window used for tBME calculation will never only contain the residual period, but will always partially contain an residual-free period. Hence, although compromise parameter sets are preferred, they will not achieve a high-density tBME value.

With $\tau = 20 \approx 2L_e$, we would have expected a flat plateau of low tBME values according to our theoretical considerations in Section 2.4. Instead, we see a slight intermediate peak of the tBME curve in Figure 9b (day 95). A possible explanation could be the wetting event at $t = 90$ days that seems to alleviate the model's difficulty to fit the data; a similar effect can be observed for the next wetting event after $t = 100$ days. Since the residuals are due to a lack of macro-pore flow behavior in the model (which would make water achieve greater depths faster), a wetting event helps to put the model into conditions that are more similar to the base-case (error free) data.

This can also be seen from the residuals plotted as Gy bars: their accumulation over time is abruptly stopped with the precipitation event at $t = 90$ days. If we look at the dip of the tBME curve, it starts decreasing at $t = 82$ and returns at $t = 111$. The total signal length is hence 29 days (111–82), which corresponds to the theoretically expected length of $\tau + L_e$.

4.1.2.2. Sensitivity of tBME to Window Size

In this test case, the misfit between the model ensemble and the data is around 0.1 m on average and spans a period of 10–15 days. The comparison of the four tBME curves reveals that a time-window size of 50%–100% residual period length is here not yet conclusive. The reason is that such a short period is not sufficient to disentangle the error signal from measurement noise (which is chosen to be on a similar order of magnitude as the residuals here) and model-internal variability due to parameter uncertainty. Also, under real-world conditions, the tBME curve for a narrow window is expected to be sensitive to short-term deviations, as daily deviations take up a larger fraction of the total residuals. For instance, a short-term variation in the system, for example, misfit during one day, can have one fifth contribution in a time window of five days. Therefore, major signals might not yet be distinguishable from noise in the tBME diagnosis with small window sizes.

We can observe that tBME with $\tau = 20$ shows a stronger signal than with $\tau = 5$ in Figure 7. This is because with τ we decide how many data points we want to calibrate on at a time. When more data points are used in the calibration per time step of the time window (larger τ is used), the more joint likelihood decreases (because the sum of squared errors in the likelihood function grows with each data point), and tBME reflects this fact.

The BME indicator strength increases with the calibration time-window length, because the task of fitting all contained data points well with any single parameter combination becomes increasingly hard (especially, since we assume uncorrelated errors in our likelihood, cf. Equation 3). Hence, wider time windows smooth out the noise, if the error duration is longer than the typical scale of noise (or unresolved, short-term processes of less importance to the modeler), and enlarges true error signals. This is what we observe here: when the time window is larger than the actual residual period (e.g., $\tau = 20$ days), the problematic periods can be clearly detected.

Overall, we take away from this first test case that a variation of τ can be very insightful and strengthens our interpretation of results.

4.1.2.3. Dynamics of Posterior Parameter Distributions

Alongside the tBME curve, we obtain a time series of posterior parameter samples for each time-window size τ. We can inspect how each parameter's posterior PDF changes with increasing or decreasing tBME over time. Figures 8 and 9 show the posterior PDFs of all five uncertain model parameters corresponding to the five illustrated time windows for window sizes $\tau = 10$ and $\tau = 20$ days, respectively. We observe the clearest trends in parameter value shifts for short window sizes τ, because parameters can be fine-tuned to very short calibration datasets, and generally do not have to satisfy varying conditions simultaneously.

For $\tau = 10$ in Figure 8, we conclude that the shape parameters α and l are most sensitive to the residual periods: During phases when dual-porosity behavior would be needed to appropriately mimic the data, α and l tend to take on higher values than during error-free (single-porosity) periods. We nicely see the error compensation attempt in those two parameters with their posterior PDFs shifting from state I to state III* (tBME returns to the high-density region due to parameter compensation) and back.

These two parameters have the effect of scaling the water retention function. Higher values correspond to lower water contents for a specified pressure head or higher pressure for a specified water content, respectively. To confirm this conclusion, we look at maximum-likelihood water retention curves obtained in the five states discussed above (Figure 10a). We also show the true water retention curves for the error-free base case and test case 1 for comparison. As expected, we see a shift of the water retention curve from state I being close to the base case toward case 1, induced by higher α and l-values that better represent the water retention function of the dual-porosity model parameterization. Hence, although the synthetically-introduced model error here is related to the model concept, the model manages to compensate the error by adjusting those parameters that determine the shape of the water retention curve.

Note that the information content of the calibration data for identifying the true water retention curve is limited, as can be seen from Figure 10a: The pressure head observations in each state (data in the five different time
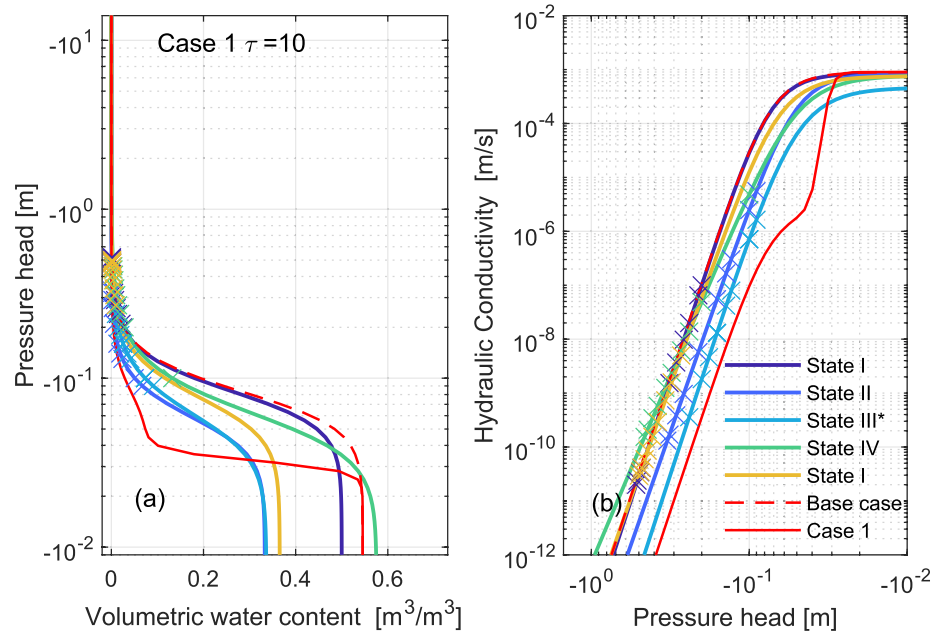
Figure 10. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 8b (synthetic case 1, window size $\tau = 10$ days). Dashed red lines show true curves of the base case; solid red lines show the curves corresponding to the perturbed dataset. x-symbols indicate the state I being close 900 to the base case toward case 1, induced by higher and l-values that better represent the water retention function of the dual-porosity model parameterization. Hence, although the synthetically-introduced model error here is related to the model concept, the model manages to compensate the error by adjusting those parameters that determine the shape of the water retention curve.

windows) are marked as crosses on the corresponding simulated curve. These data mostly cover the dry range, where the water retention curve becomes a vertical line. None of the pressure head observations in that range is informative, because differences are very subtle and hence the true water retention curve cannot be identified. This is a particular result of the highly porous volcanic soils present in our case study region (Wöhling et al., 2008). In these soils (similar to sand), the water content drops dramatically over just a narrow range in matric potential.

Hydraulic conductivity as a function of pressure head (shown in Figure 10b) is dominated by the parameter compensation effect that causes the curves to shift from state I closest to the base case toward the curve of case 1 for states II and III*. The unsaturated conductivity of in case 1 is much lower than in the base case, leading to a slower movement of infiltration fronts and larger pressure heads.

As expected, parameter compensation effects become less dominant when the window size is increased beyond the residual period length (illustrated for $\tau = 20$ in Figure 9 and Figure A2 in the Supporting Information S1). Parameters now have to find an optimal compromise solution whenever (parts of) the error period are contained in the calibration time window, and this reduces the shifts in PDFs and water retention curves or unsaturated hydraulic conductivity functions between time window positions. Yet, differences between the individual parameters can still be found, which can help gain further understanding of how these parameters act in the model and how they possibly compensate for model error.

4.1.3. Case 2: tBME Curve Given Forcing Error

The second case illustrates the analysis of tBME curves in the presence of forcing errors. Figure 11 shows the forcing (precipitation), the manually perturbed dataset, the error-free base case dataset for comparison, and the resulting tBME curves together with their sampling distributions for the four window sizes τ. Recall that, to construct the dataset for this test case, we have run HYDRUS-1D with the base case setup. Only precipitation was set to zero during the pre-defined error periods, which are marked with red-shaded boxes in Figure 11a.
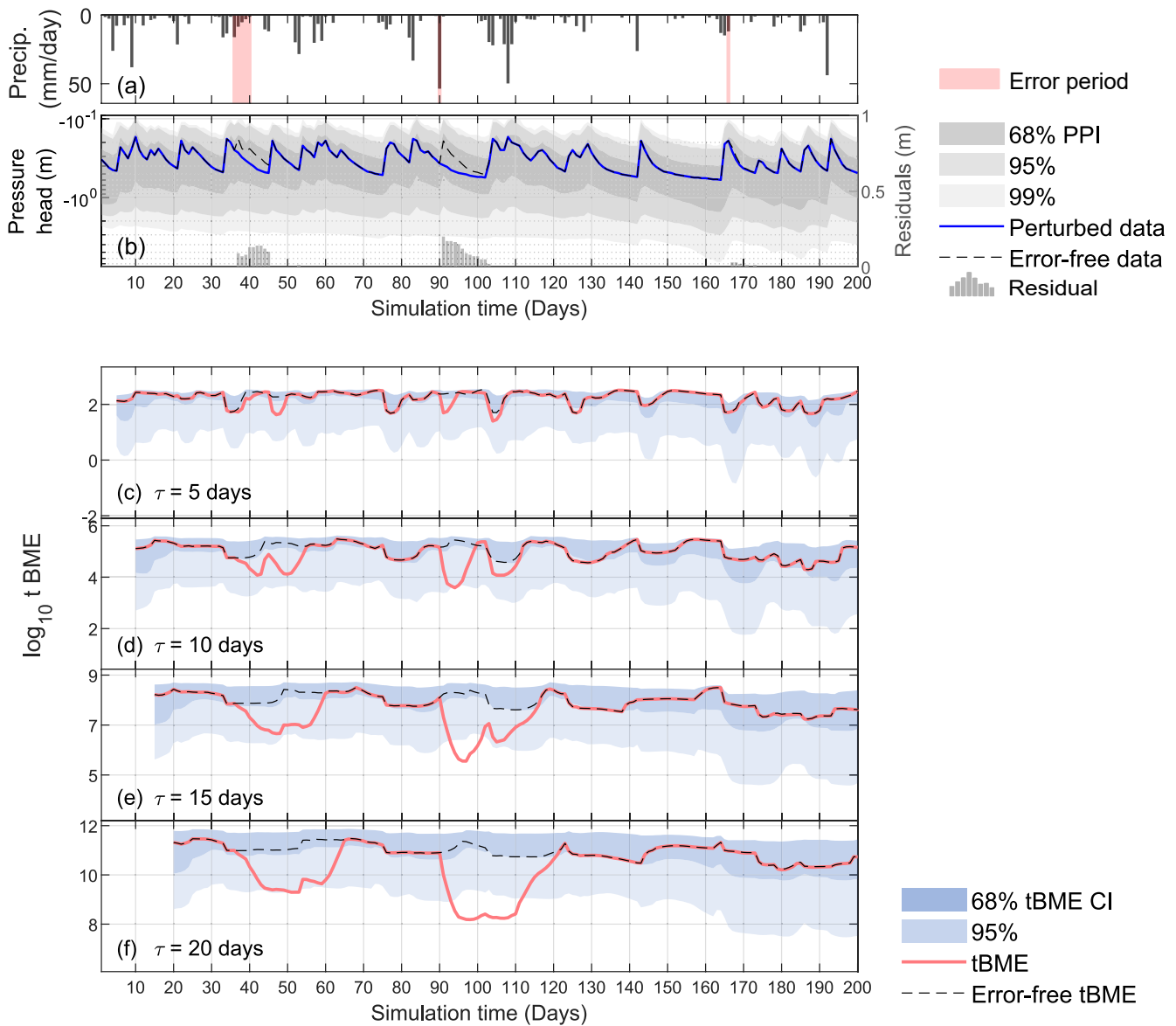
Figure 11. tBME curves given forcing error (synthetic case 2: Precipitation marked in red has not happened at the field site, but is used as input for our model).

4.1.3.1. Error Detection

It becomes apparent that only the first two error periods lead to a significant accumulation of residuals between perturbed data and the model's base case realization. Both periods are clearly detected by the tBME indicator at window sizes larger than five days, with tBME leaving its high-density region and, for the error period with largest residuals, even the complete range of the sampling distribution.

The theoretically expected behavior of the tBME curve can be seen beautifully at the example of $\tau = 10 < L_e$ in Figure 12c. The fall and rise of tBME takes approximately $\tau/2 = 5$ days, and the intermediate plateau of high tBME values persists for exactly $L_e - \tau = 3$ days. The second dive is somewhat distorted, with a fast fall and a slow rise. This effect might be due to the precipitation events that hit the system right in this "recovery" period. While the second error triggers a similar tBME signal, the third error period is both short and weak (only a small precipitation event is set to zero), hence it shows no visible signal.

Table 4

Dual-Porosity Mualem-van Genuchten Parameter Values Applied to Perturb the Base-case Dataset During Error Periods in the case of Model Structural Error (case 1); Parameters for the First Pore Subsystem Are Identical to Those Given in Table 2 (Base case), With $w_1 = 1 - w_2$

	w_2	α_2	n_2
Layers 1 and 2	0.85	31	15

Table 5

Mualem-van Genuchten Parameter Values Applied to Perturb the Base-case Dataset During Error Periods in the case of Superimposed Errors (case 3), Where w_2, α_2, and n_2 Are Switched on Only During Error Periods Defined in case 1

	θ_s	α	N	K_{sat}	L	w_2	α_2	n_2
Layer 1	0.4146	6.1609	2.5970	0.0001	0.4049	0.2	31	15
Layer 2	0.3860	2.2569	5.5489	0.0004	0.3108	0.2	31	15
Layer 3	0.5719	2.9203	4.5101	0.0008	0.9006	0	31	15

4.1.3.2. Sensitivity of tBME to Window Size

As discussed in the context of case 1 (model-structural error detection), moving toward longer time windows increases true error signals and flattens out noise and the fluctuations of the tBME curve that are caused by parameter variability within the model. Further results are shown in the Supporting Information S1.

Based on the two investigated test cases so far, we conclude that the selection of an appropriate window size τ depends on many factors, such as magnitude of model residuals, level of measurement noise, duration of residual occurrence, and internal variability of the model ensemble. Although many of these factors might be unknown in practice, it helps if the initial τ selection is based on (expert) knowledge of the expected error and its duration, which hydrologists or generally modelers typically have some ideas about. If expert knowledge is not available, we recommend analyzing tBME curves with a prior selection of window sizes to assist model error detection and and to provide insights about temporal scales of model error occurrence.
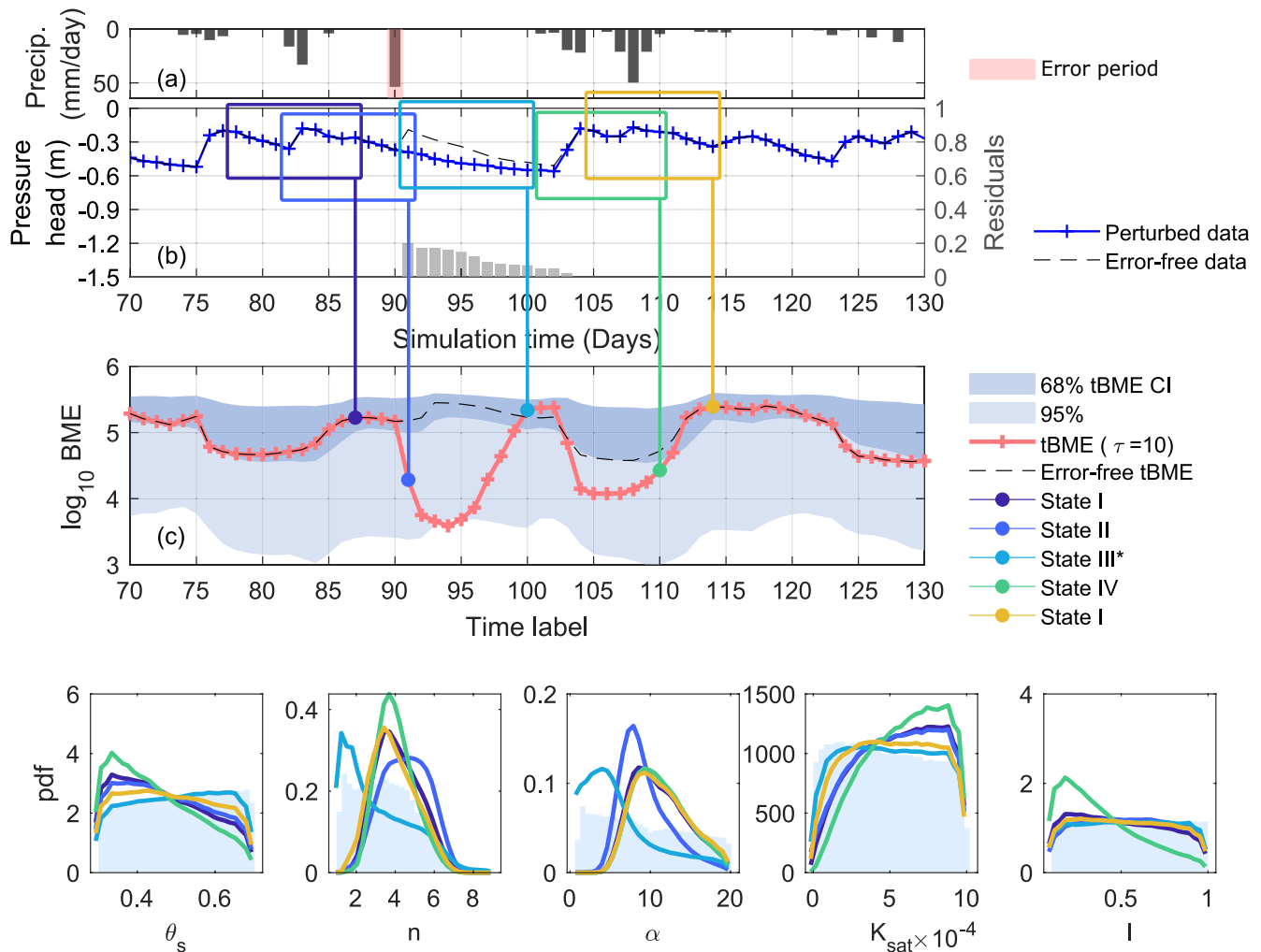
Figure 12. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 10$ (Synthetic case 2).
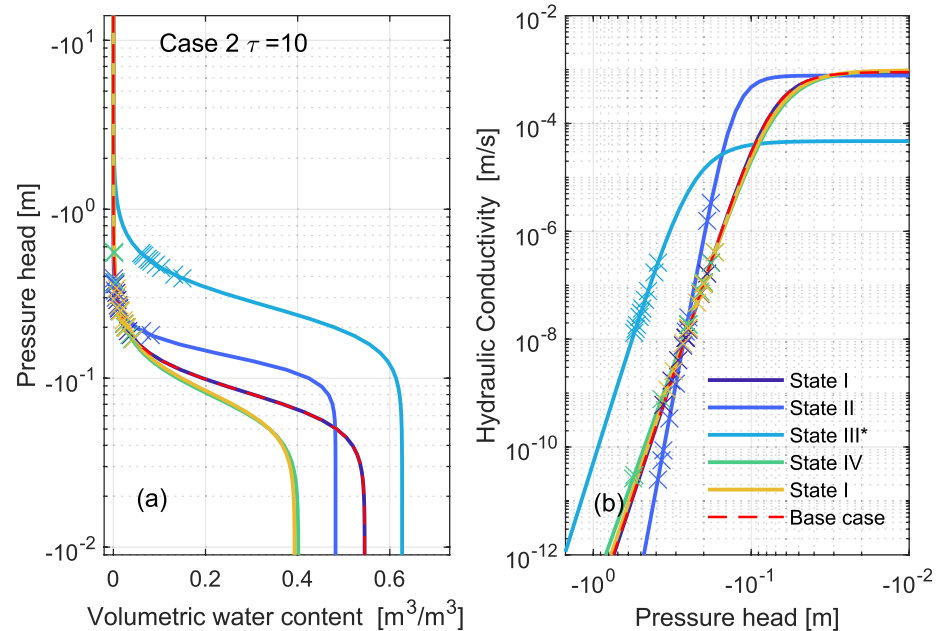
Figure 13. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 12b (synthetic case 2, window size $\tau = 10$ days). Dashed red lines show true curves of the base case; there are no "true" curves for the perturbed dataset because forcing error does not impact the soil hydraulic characteristics.

4.1.3.3. Dynamics of Posterior Parameter Distributions

We find that, also in the case of forcing error, a compensation effect is achieved by shifting parameter distributions. This can be seen from the climb of the tBME curve for $\tau = 10$ days at around $t = 100$ days, and the second drop right after (Figure 11d). Around this point in time, the time window almost exclusively captures the residual period, and parameter values are reweighted in Bayesian updating to achieve an optimal fit of the perturbed data. A rise of the tBME curve during a residual period can only be explained by shifting parameter values, and this explanation is confirmed by inspecting the dynamic posterior parameter distributions for the selected windows of size $\tau = 10$ days in Figure 12.

The MVG shape parameters α and n seem to be responsible for error compensation: both take on distinct low values only during the residual period. Inspecting the maximum-likelihood water retention curves obtained for the five states (Figure 13a), we find a clear shift toward higher water content at a given pressure head during residual periods. This is the only possible way for the model to fit the observed pressure head data despite the surplus of precipitation in the soil. It is achieved by the interplay of low α, low n and high θ_s. Similarly, hydraulic conductivity is shifted to higher values at a given pressure head (Figure 13b). This leads to faster drainage of the soil profile, which there in turn will result in lower θ and higher h. This is clearly a compensation effect.

To find out whether these successfully compensating parameter sets are clearly identifiable from calibration within the short window of 10 days, we additionally show the 95% posterior credible interval of the water retention curves and the unsaturated conductivity functions of states I and III* in Figure 14. The intervals of both states are extremely large (much wider than in case 1, cf. Figure A3 in the Supporting Information S1) and allow for a wide range soil hydraulic functions. This warns us that there is no dominant pattern of error compensation - which is plausible since model error is caused by model forcing in this test case, not by modifications of the model parameterization.

Sliding out of the residual period, it seems that K_{sat} and θ_s are now taking care of the excess water in the model (since it had to cope with precipitation input that is missing in the true system). K_{sat} is clearly increased during this phase (and rather insensitive else), and θ_s is reduced (cf. posterior PDFs in Figure 12). The model seems to struggle when going back into the residual-free period, because what happens in the perturbed data is that the current saturation is taken as initial value to continue calculations with the original parameter setup. But the model is in
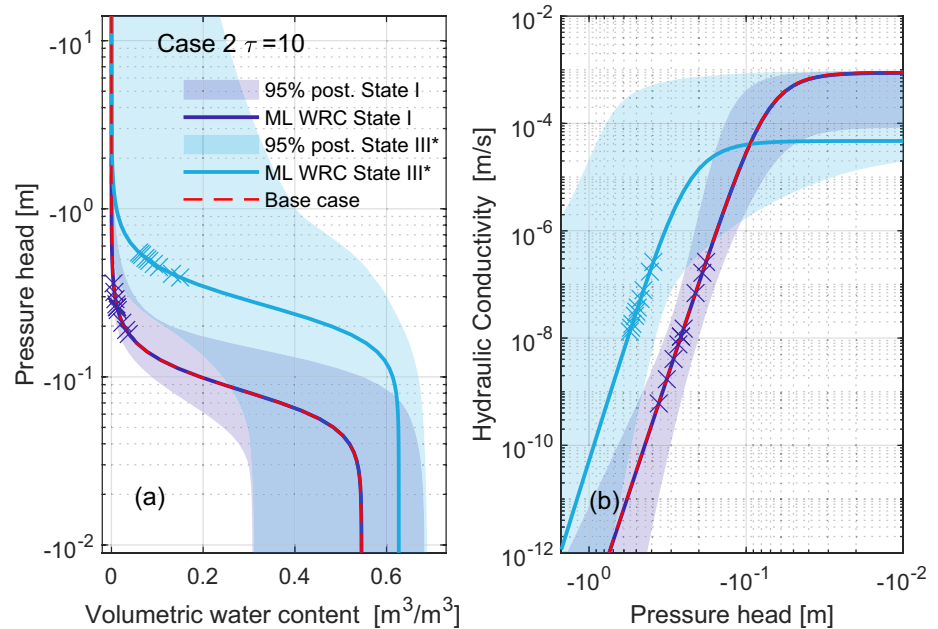
Figure 14. Posterior credible intervals of water retention curves (left) and unsaturated hydraulic conductivity curves (right) of the top soil layer for states I and III* (synthetic case 2, window size τ = 10 days).

a different state because it tried to compensate for excess water and hence carries some "history" with it. On top, another heavy precipitation event is hitting the catchment during the recovery phase after $t = 100$ days. One might expect that this event should equalize the tensiometer traces; but since the model is not doing very well during infiltration events anyway, the original dip in the reference BME curve is enhanced. All parameters are shifting in a way distinct from all other phases (green time window and PDFs in Figure 12). Both effects combined might explain why pre- and after residual-period parameter values are not the same for θ_s and K_{sat}, and why water retention curves look different between state I before the residual period and state I after. But, importantly, $K(h)$ has gone back to the same state I as before.

We conclude that inspecting the dynamic posterior parameter PDFs, posterior water retention curves and posterior unsaturated hydraulic conductivity curves including their remaining uncertainty can help put the detected model error into context. The sensitivity of individual parameters to the sliding calibration time window can provide information about the type and source of error.

4.1.4. Case 3: tBME Curve Given Superimposed Errors in Model Structure and Forcing

In reality, model errors occur at an unexpected frequency and at various temporal scales, so multiple residual periods are expected to overlap each other. The analysis of multiple superimposed errors is demonstrated with the setup of case 3 described in Section 3.4.

Figure 15a records the forcing (precipitation); Figure 15b shows the manually perturbed dataset and the error-free base-case dataset for comparison. The red-shaded boxes in Figure 15b mark the periods when the model suffers from structural error (reality is in dual-porosity mode), and the red-shaded boxes in Figure 15a show the periods when the model suffers from erroneous forcing. The model misfits caused by these two errors partially cancel out. The net residuals are presented by the Gy bars in Figure 15b. The resulting tBME curves are plotted together with their sampling distributions for the four window sizes τ in Figures 15c–15f.

4.1.4.1. Error Detection Through Varied Window Size

Consistent with the previous two cases, we do not perceive clear signals from the shortest window size $\tau = 5$ days. From $\tau = 10$ days on, we see strong signals during all three residual periods; and these signals are massively intensified when increasing the window size toward $\tau = 20$ days.
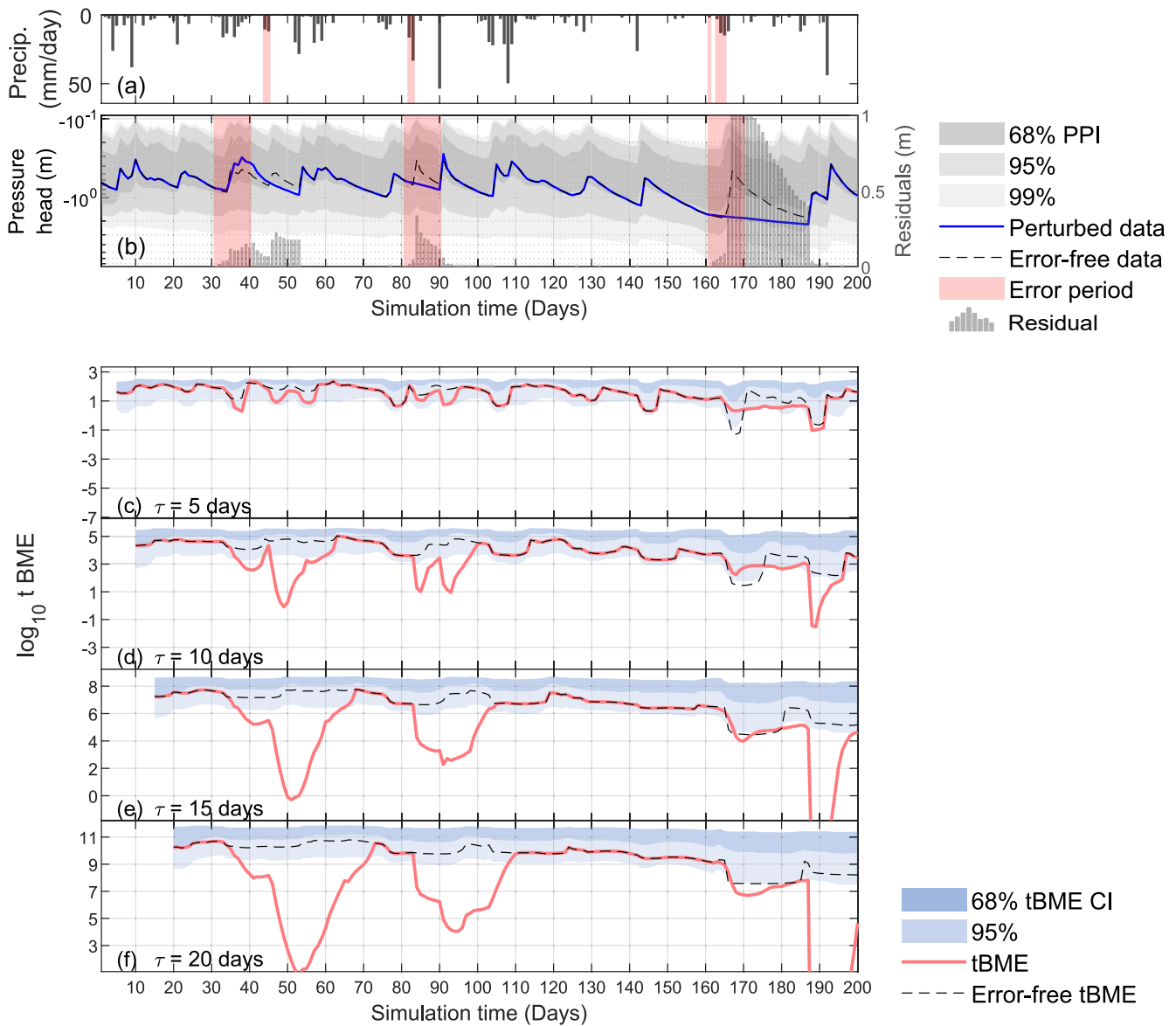
Figure 15. Results of tBME analysis if model structural and forcing errors are superimposed (synthetic case 3).

Recall that the dotted lines represent the tBME curves of the chosen model realization before it was perturbed to generate the synthetic data set for this test case 3 (labeled with "error-free tBME" in Figures 15c–15f).

Note that this dataset is chosen to be a different one than the base-case dataset underlying cases 1 and 2; with the motivation to now move as closely to realistic dynamics as possible. The realization used here does not fall within the high-density region of the prior predictive distribution (cf. e.g., days 160–170 in Figure 15b), so short segments of the error-free tBME curves fall outside the shown intervals of the BME sampling distribution (e.g., days 160–170 in Figure 15c and 15days). Note that, per construction, it must fall inside the sampling distribution, just at a very extreme tail. However, we observe that this type of signal is not enlarged when increasing the time-window size. This finding again supports our faith in the method: true model errors will be detected by varied window sizes through drastically intensified signals; model behavior within the statistically plausible range will not produce a strong signal.

The superposition of errors can be seen when comparing the tBME curves for $\tau = 10$ days with the tBME curves for larger window sizes during days 30–70: A time window of 10 days exactly separates the two residual periods (the residual caused by structural error first, the residual caused by precipitation input error second); at larger time
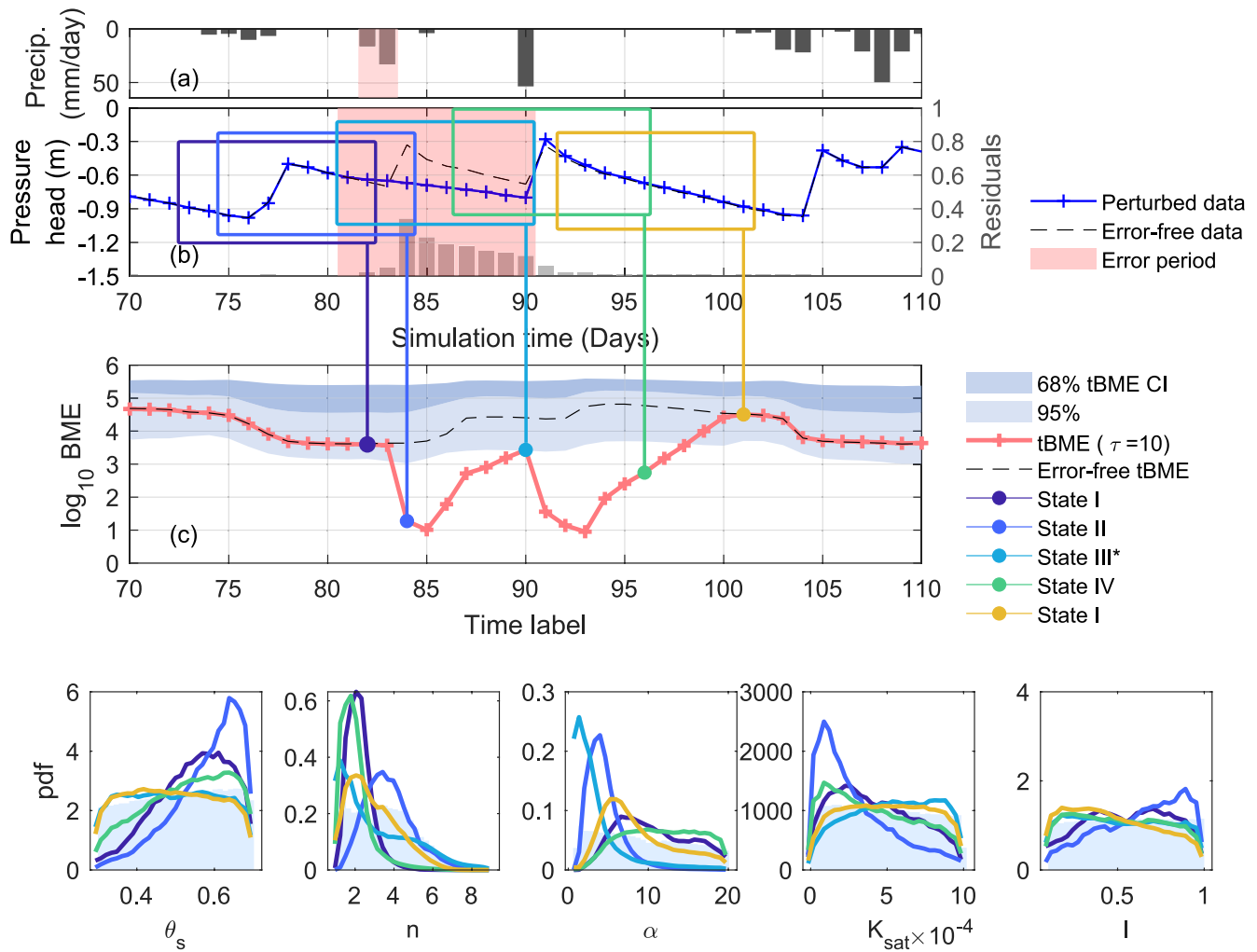
Figure 16. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 10$ (synthetic case 3).

windows, these two signals triggered by the two residuals merge. This points us effectively to the time scales of both errors, which are about 10 days of large residuals each.

When one residual period overlaps the other (days 80–90), we see a double peak for $\tau = 10$ days, and a merged signal for larger window sizes. Inspecting the double peak in more detail in Figure 16 reveals that the rise of the tBME curve after its first dive into the residual period comes too early - theoretically, we would expect it at $\tau/2$ days after the onset of the signal because $\tau \approx L_e$ (cf. Section 2.4); that is, at $t = 88$ days. Here, instead, we see a rise already at $t = 86$ days. This is due to the fact that the superimposed errors cancel out to some degree, and hence, tBME has a chance to rise again even earlier. As expected, it arrives at its temporary peak at L_e days after the onset of the signal (because then, the residual period is completely captured by the window), here at $t = 90$ days. Sliding out of the residual period again leads to a decline of tBME, which should arrive at its intermediate minimum after $\tau/2$ days, but again is too early here. This is why the rise to the end of the error signal appears very long.

The steep decline and slow return of the tBME curve is indeed the very characteristic of superimposed, counteracting errors: the partial canceling of residuals affects the originally expected tBME behavior for individual errors in that way. The effect becomes even more clear if we carefully think about which timing to expect here for an ideal case of a single error signal: since τ is in fact slightly larger than L_e, the two falls of the tBME curve should last $\tau/2$ days, and the two rises should last $L_e - \tau/2$ days; that is, the fall would theoretically take longer
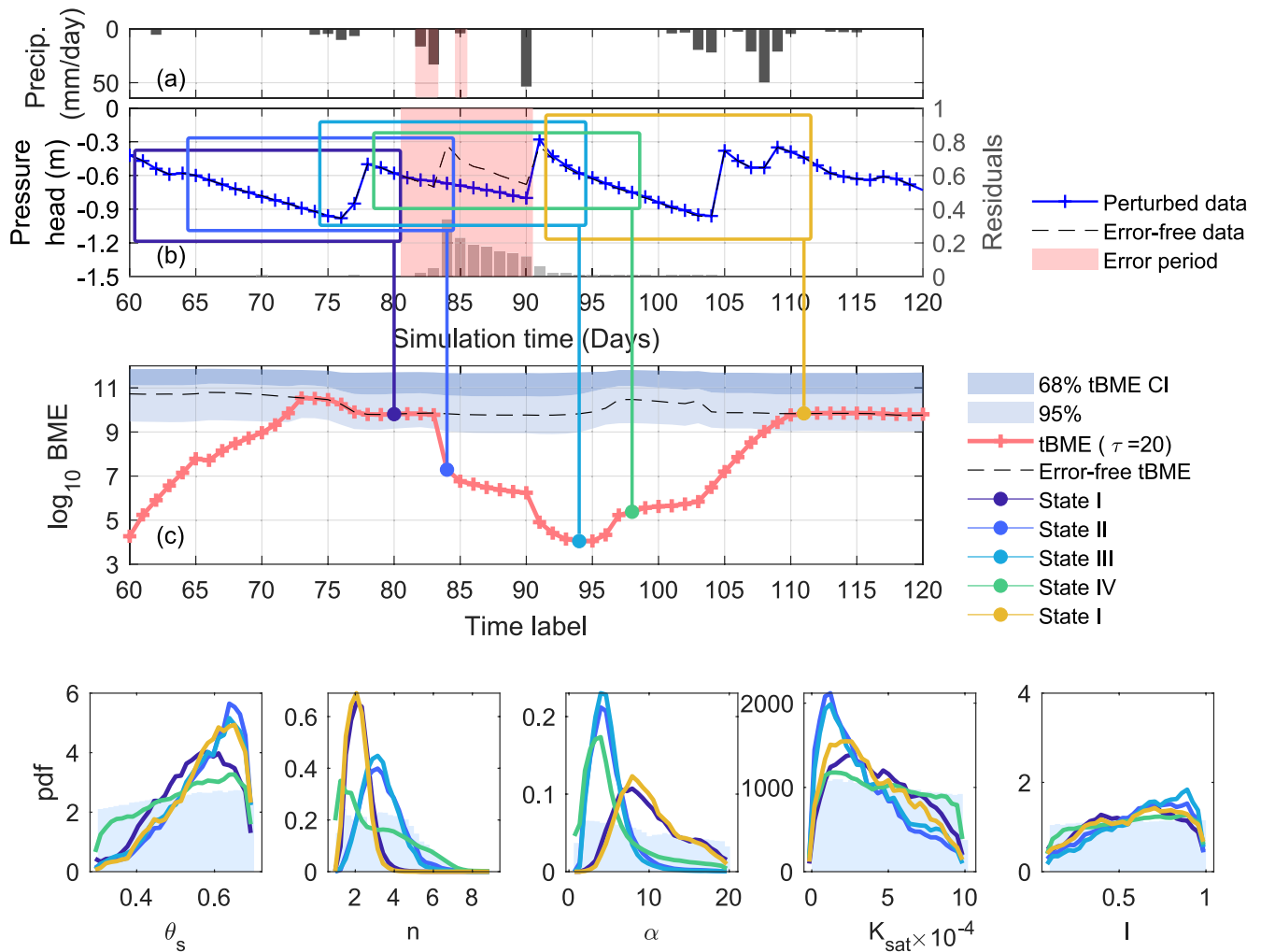
Figure 17. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 20$ (synthetic case 3).

(five days) than the rise (two days) to a plateau of three days length. What we see is the opposite behavior, and this is due to the counteracting effect of the error sources.

Also, the merged signals for larger τ in both residual periods (days 30–70 and 80–90) look distinct and can be distinguished from the case of parameter compensation as seen in cases 1 and 2. In the case of parameter compensation during a single-error period, we observed a stable plateau in the tBME curve when increasing the window size. Here, we see distinct steep drops from such a plateau. Similar to our interpretation above for $\tau = 10$ days, we find that the fall to the minimum plateau, expected to occur for $t = 91$–103, seems to be dampened ($\tau = 20$ days, Figure 17). The plateau is interrupted by an additional dive, before the tBME starts to climb back up as expected from $t = 104$ days on.

We found that determining the residual period length as total signal length L_s minus τ yielded seven days, which only identifies the period of large superimposed residuals. Hence, for directly superimposed residuals, inferring the individual time scales will be hard (at least for the less pronounced residual), but varying the time-window size together with low-noise data could help identify also this one.

4.1.4.2. Dynamics of Posterior Parameter Distributions

When inspecting the dynamic posterior parameter distributions for this test case 3 at a window size of 10 days (Figure 16), we see attempts to parameter compensation that seem disrupted: judging from the tBME value during state III, a parameter shift can accomplish a return to the edge of the high-density region of the tBME
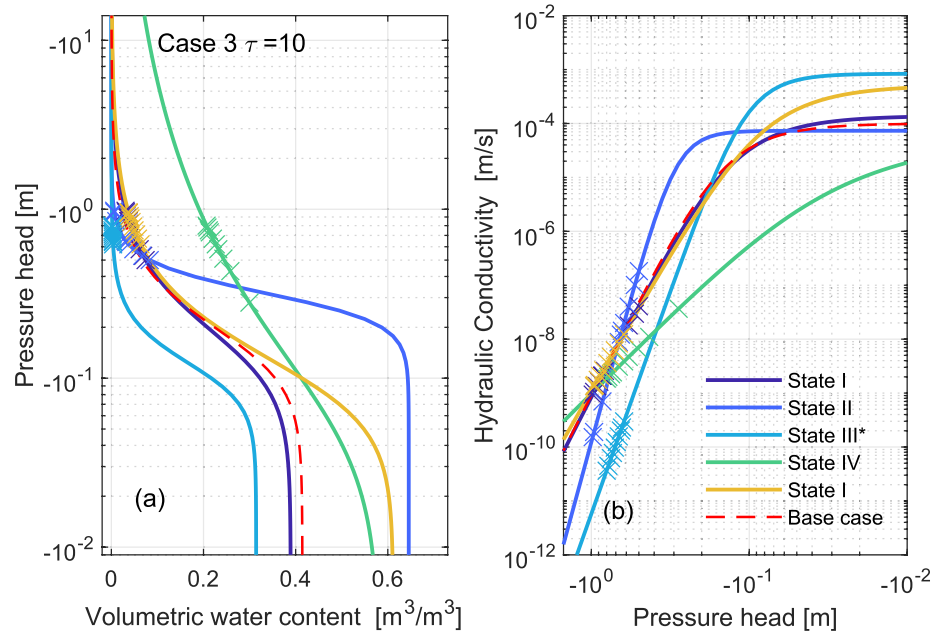
Figure 18. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 16b (synthetic case 3, window size $\tau = 10$ days). Dashed red lines show true curves of the base case; again, there are no "true" curves for the perturbed dataset.

sampling distribution; however, this shift is not consistent throughout the residual period. For all parameters but α, the first reaction of the parameter distributions during state II produces much more pronounced shifts away from state I (residual-free period) than during state III. This again confirms a "mitigating" effect of error superposition in this case. The role of α has indeed not changed - it takes on very small values during error compensation as in test case 2 (erroneous forcing), but the interplay with n seems different here.

We further inspect the maximum-likelihood water retention curves and unsaturated hydraulic conductivity functions for the five states in Figure 18 and in Figure 19. Also, here, we see mixed compensation signals from the two previous cases.

The fact that α clearly shifts to counteract the forcing error, but not the structural error (α moved to higher values in test case 1), combined with the much larger shift of the water retention curve toward higher water contents as in case 2, suggests that the forcing error is playing the dominant role here. Such interpretations thus help identifying the source and role of model errors. For interpreting real-world analysis results (as to be done in Section 4.2), we can thus recommend to create synthetic scenarios with hypothesized errors to investigate the corresponding outcomes of tBME, posterior parameter distributions and posterior soil hydraulic functions for comparison.

For $\tau = 20$ days, we see more consistent parameter compensation attempts in both marginal parameter PDFs (Figure 17) and soil hydraulic functions (Figure A10). "Averaging" of the impact of both errors seems to produce clearer compromise solutions in parameter values. α is decreased during the residual period, n is increased during the residual period. Again, this represents a mitigation of the forcing error, not so much of the structural error. θ_s and K_{sat} show trends of increased saturated water content and decreased conductivity during (and after) the residual period, which is similar to our findings from case 1 (structural error), and not case 2 (forcing error).

Note that the error timing in case 3 is different and does not allow for a detailed 1-to-1 comparison: while the structural error period in this third test case is the same as in case 1, the forcing error has been chosen to overlap, so it is earlier in case 3 than in case 1. This leads to a severe precipitation event hitting the catchment during its error recovery phase in case 3; in case 2, this event was subject to forcing error inside the error period. Hence, both scenarios naturally request for different strategies to bring the model back into a state that matches the original error-free data set well.
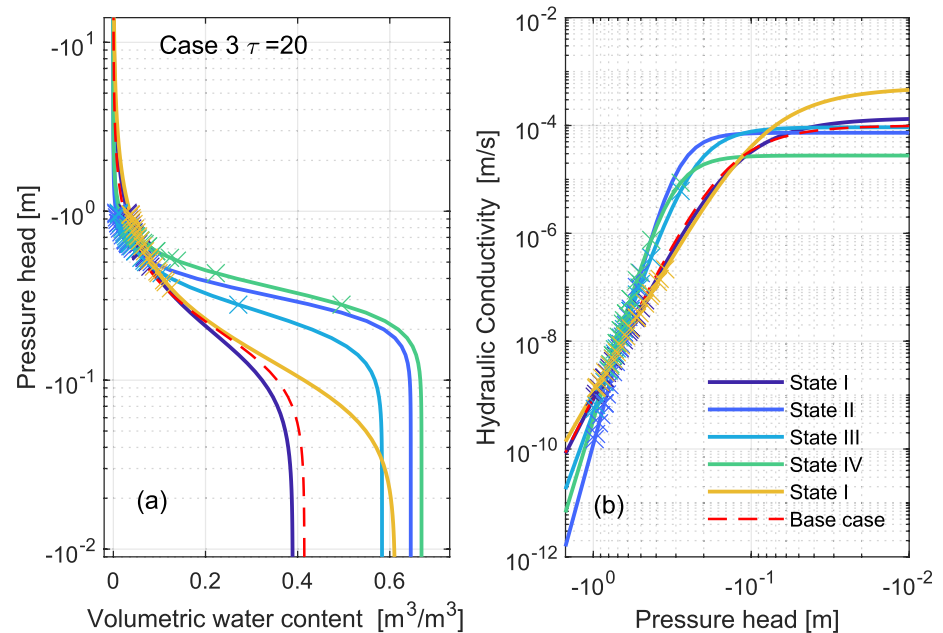
Figure 19. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 12b (synthetic case 3, window size $\tau = 20$ days). Dashed red lines show true curves of the base case; there are no "true" curves for the perturbed dataset because forcing error does not impact the soil hydraulic characteristics.

4.1.5. Summary of Findings From Synthetic Test Cases

We found that tBME curves can be quite wiggly due to parameter uncertainty or due to other factors that increase the dynamic and variability in ensemble predictions. Also, the tBME sampling distribution can show dynamics over time. In conclusion, no matter how wiggly the tBME curve obtained for real data, if it stays within the tBME sampling range, there is no substantial (statistically significant) reason to believe in model error.

We do not give any general recommendations on a significance threshold for "low tBME" as an indicator for error occurrence; rather, we encourage the modeler to test increasing window sizes and identify those periods that consistently produce clearer signals. For more confidence, the modeler could perform synthetic tests as demonstrated here to learn about the model's tBME reaction to a certain level of error under idealized conditions.

Results have shown that different types of errors and even superimposed errors can be detected with our proposed method. The interpretation of the tBME behavior is guided by the theoretical considerations in Section 2.4. We can, for example, easily identify the residual period length from visual inspection of "tBME valleys." We found that the tBME curves in our test cases mostly followed the theoretically expected behavior closely; however, changing conditions (such as in forcing) make the tBME curve leave its theoretical trajectory. Also, superimposed errors that tend to cancel out have a specific impact on the tBME curve: they produce a steep decline and a slow return. Hence, by comparing the actually obtained tBME curve with the theoretical sketches and by paying attention to other factors that could change the model ensemble's behavior before and after an error period, we can extract valuable information about the type and duration of model error.

Analyzing of dynamic shifts in time-windowed posterior parameter distributions allows to gain an understanding of how parameters possibly compensate for model error. These insights may, in turn, provide avenues for model refinement. In the specific case of soil hydraulic modeling, interpreting posterior water retention curves and unsaturated hydraulic conductivity functions proves highly valuable. In the considered setup of our case study, the range of "synthetically observed" data values covered only a very small (dry) portion of the full saturation range, and hence, data are only slightly informative for identifying most plausible soil hydraulic parameter values. Although this is a well-known and important phenomenon, it is often forgotten during the analysis of hydrological model results.

Potential users of our proposed method might ask themselves whether just looking at residuals would be sufficient to obtain a model error signal, avoiding the hassle to go through Bayesian updating altogether? Unfortunately, inspection of residuals only would not do the trick: remember that what we showed as the Gy bars in Figures 7–17 are the residuals between the error-free data set and its perturbed version - this is unknown in practice. We would have to look at average residuals of ensemble predictions versus data instead, or at best-fit-residuals, and both can be misleading if parameter compensation occurs and/or superimposed errors tend to cancel out. Thus, it is beneficial to make use of the Bayesian framework: It allows error detection via tBME and deeper interpretation with respect to error compensation based on dynamic parameter posteriors.

4.2. Results of tBME Analysis for Real Observation Data

We now use the actually observed Spydia data (Section 3.2) to construct the tBME curves and wish to identify model error in our tested ensemble. The dataset has not been perturbed with extra errors. Note that the overall time period of the calibration dataset and the simulation is the same as in all synthetic cases; hence, we can use our findings from the base case and the test cases 1–3 to help us interpret the results from this real case - since the source of error is obviously unknown now. This case is a demonstration of the potential usefulness of our method under realistic conditions, while the synthetic cases have served as a validation of our approach.

Figure 20a visualizes the model forcing (recorded precipitation); Figure 20b shows the real dataset and the prior predictive range of our model. Error periods and residuals are obviously unknown now. The obtained tBME curves are plotted together with their sampling distributions for the four window sizes τ in Figures 20c–20f.

4.2.1. Error Detection Through Varied Window Size

The drops in the tBME curve of $\tau = 5$ indicate multiple potential error occurrences, which could, however, be clouded by measurement noise as noted in the synthetic test cases. Compared with the observations, we notice that drops in the tBME curve always occur when a recession phase abruptly changes to a wetting event, that is, when the soil water pressure jumps to higher moisture contents. This implies that the model cannot describe highly dynamic responses well, and consequently tBME goes down.

With the wiggly nature of the tBME curve under real-world conditions, the sampling distribution proves to be very helpful for interpretation. That is also why we designed synthetic case 3 such that it mimics real-world dynamics, to make ourselves familiar with interpretation of such curves.

To distinguish between statistically significant errors and noise, we search for enlarged signals along the vertical axis of Figure 20 (tBME curves of increasing τ). From the tBME curve for $\tau = 20$, we observe amplified signals that fall out of the tBME sampling distribution during days 34–53, 75–96, and 105–128, which we consider as error signals. The interval between days 130 and 160 does not show noticeable model errors, since the tBME curve does not shift downward significantly with increasing window size.

The tBME curves of larger window sizes do not indicate well-separated single peaks, which suggests that either the time scale of model error is on the order of 10 days or less; or model error events are overlapping and hence leading to a smeared-out signal. Both hypotheses are probably true if we assume that model error is related to the time of wetting (one to three days): with increasing τ, often the next event is already included in the window.

In Figure 21, we can inspect the dive of the tBME curve between days 103 and 118 in more detail. The model clearly leaves the high-density tBME trajectory for a period of $L_s = 14$ days. Recalling the theoretical properties of the tBME indicator, the underlying error period (time period with significant residuals) should be of length $L_s - \tau = 4$ days. Further, the pattern of the tBME curve looks more similar to superimposed errors (cf. synthetic case 3) than to individual errors (cf. cases 1 and 2). Again, superimposed, short-term errors seem plausible here if we assume that they are related to soil wetting, given that several days of precipitation with varied intensity occur during this period.

For simulation days 160–170, we perceive a signal that has a very clear, familiar shape, with a widening valley for inc reasing time-window size. This signal formally does not leave the reference interval obtained by sampling tBME from the model; however, the sampling range is here very wide due to the increased model-interval variability. We therefore consider this to be a sign of model error - during this period, the model structure itself may
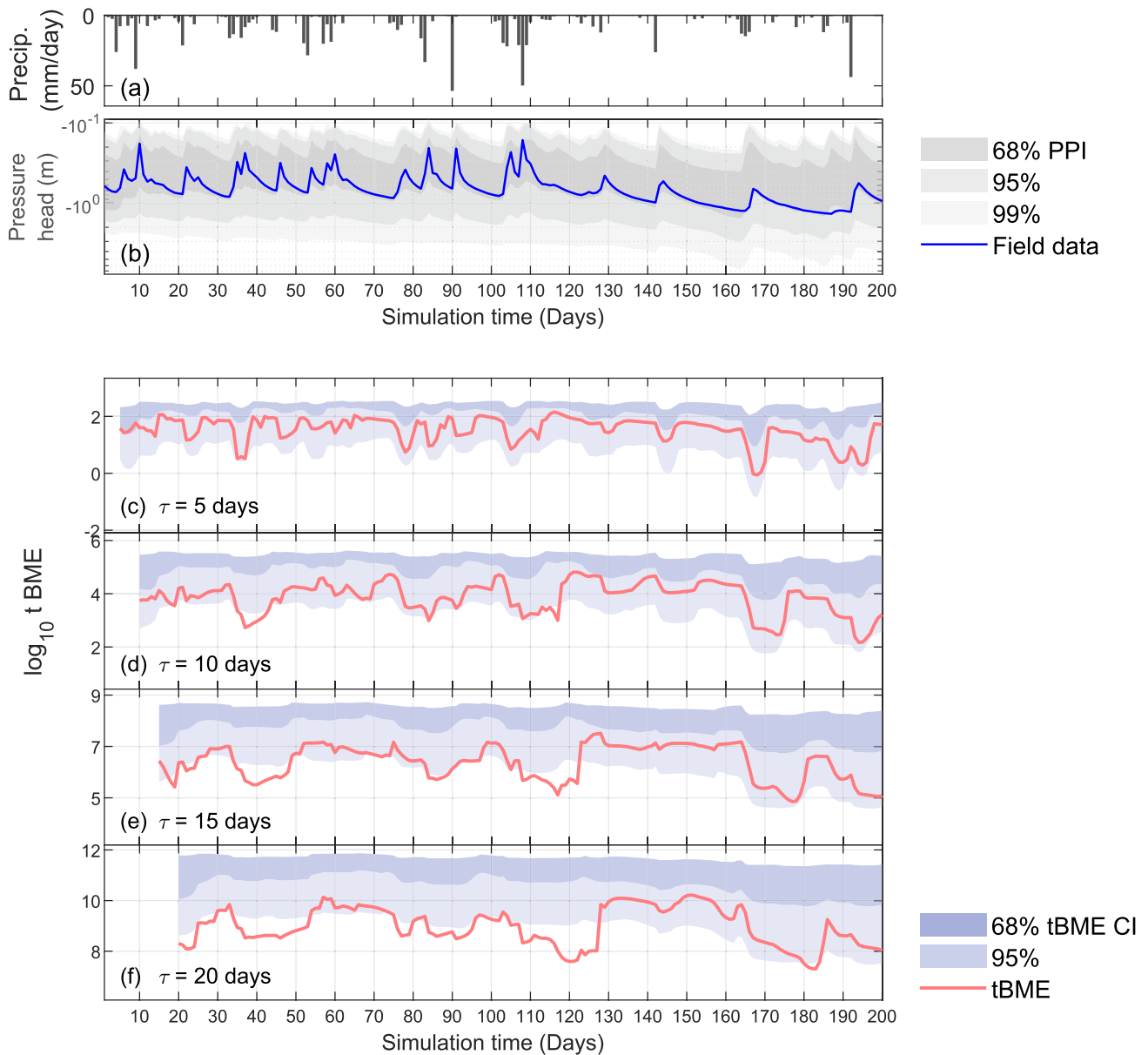
Figure 20. Results of tBME analysis for real field data.

not be disqualified as "quasi-true," but most of the parameter ensemble is. Certainly, the model is struggling here, and reasons should be investigated.

4.2.2. Dynamics of Posterior Parameter Distributions

Inspecting the dynamic posterior parameter distributions for $\tau = 10$ days in Figure 21, we see clear attempts to parameter compensation: the MVG shape parameters α and n both are increased during the low-tBME period, and the pore connectivity l is decreased. Additionally, low values of saturated water content θ_s are preferred, paired with higher conductivity K_{sat}. This reaction is similar to the pattern of parameter compensation that we have seen in test case 1 (structural error, missing dual-porosity component in the model).

Yet, judging from the tBME curve, we conclude that these compensation attempts through shifting parameter values are not highly successful. They seem to dampen the fall of the tBME curve to some degree, but they cannot bring its value up to the higher-density region of the model. This underlines that the model most likely suffers from structural error related to the wetting phases (e.g., missing macropore flow, surface runoff, hydrophobicity)
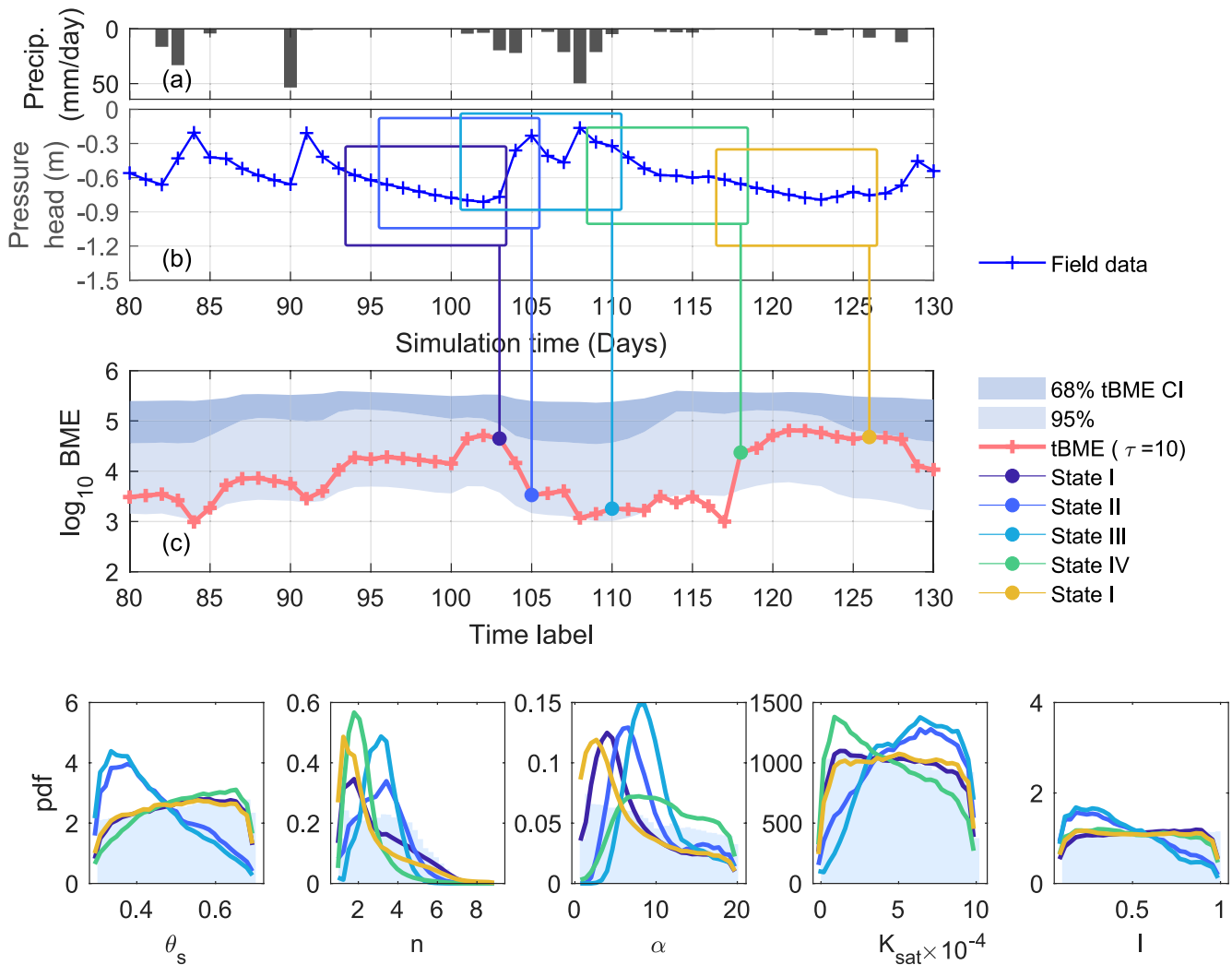
Figure 21. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 10$ (real observation data).

that needs to be treated by modifying the model structure; time-dependent parameters could here only help to a limited extent.

When inspecting the dynamic water retention curves and unsaturated hydraulic conductivity curves in Figure 22, we find that their shape in most states does not resemble the typical shape that our model ensemble favors. This supports the hypothesis that an error dominates which is not related to parameter misspecification. The remaining uncertainty after calibration on this short time window is large (see Figure 23), but still the maximum-likelihood curve of state I falls outside of the 95% credible interval. This emphasizes the struggle of the model to fit the data well, even during periods with a high-level tBME value.

The discussed findings are confirmed when inspecting the posterior parameter distributions for $\tau = 20$ days in Figure 24. For this window size, we see a total signal length of about 24 days, which yields the same suggested residual period length of 4 days. Also, the steep return to the higher-density region of our model is at the same relative position and suggests that the model is coping much better with the period after $t = 109$ days. This exactly corresponds to the transition between heavy precipitation and drier conditions, and supports our hypothesis that the model is much better at representing soil water redistribution (soil drying) than at simulating infiltration events (soil wetting).

Within these longer time windows, α and K_{sat} seem to try and compensate heavily during the recession phase of the precipitation event at $t = 108$ days. The other parameter distributions look similar to those for $\tau = 10$ days.
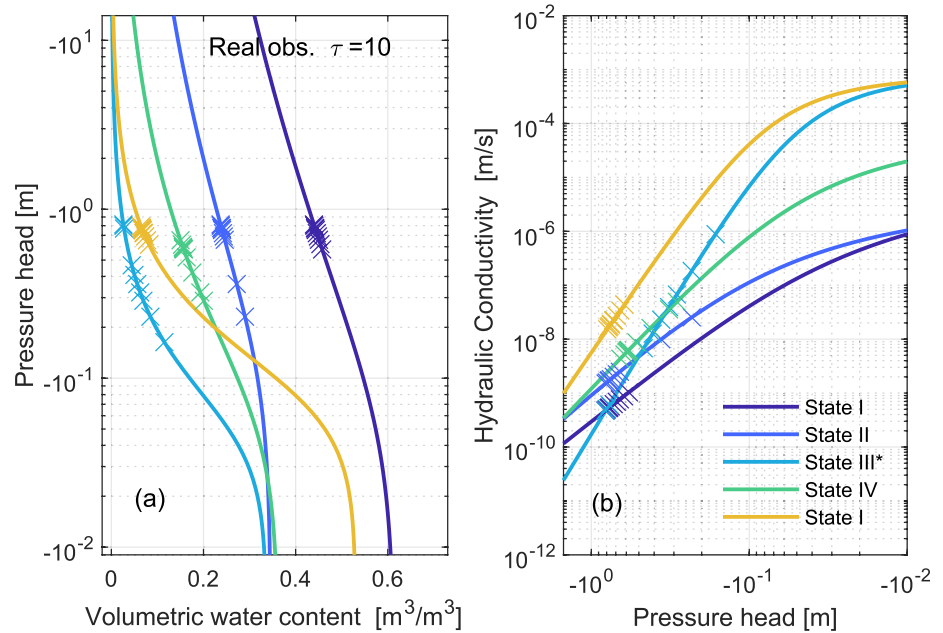
Figure 22. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 21b (real observation data, window size $\tau = 10$ days).

This suggests that averaging over longer periods of time does not require opposite directions of parameter compensation, and hence indicates potentially superimposed errors of the same type. This again supports our hypothesis of errors of our model related to infiltration processes. To improve this model, we could allow the macropore flow fraction w_2 as another uncertain parameter, and see whether it could effectively compensate most of the detected model error. Given the current error patterns revealed by our analysis, we would expect that w_2 would be larger than zero during the residual period; with the effective soil subsystem, the model misfit would be improved. However, this attempt at actual model improvement is beyond the scope of the current study.
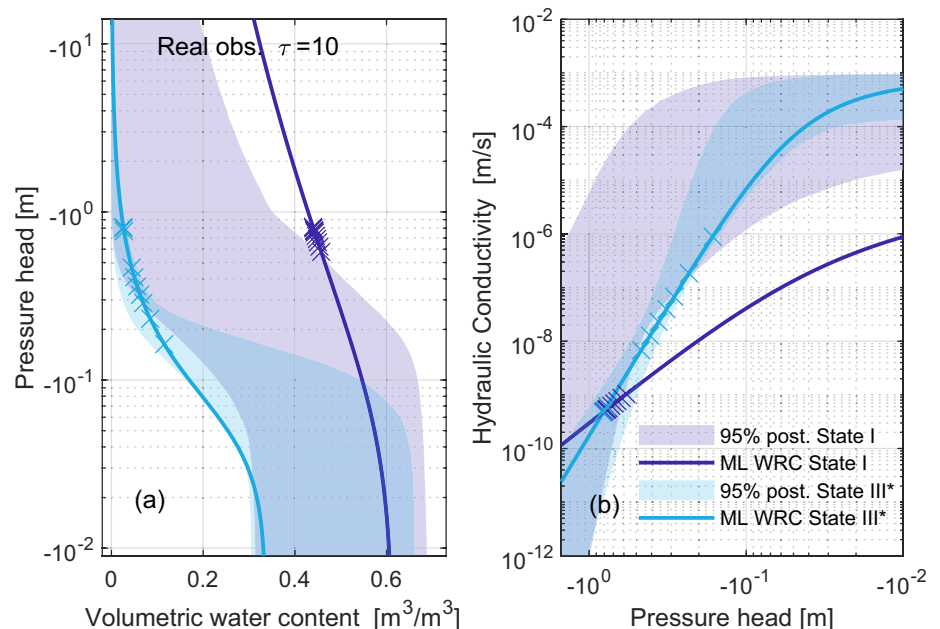
Figure 23. Posterior credible intervals of water retention curves (left) and unsaturated hydraulic conductivity curves (right) of the top soil layer for states I and III* (real observation data, window size $\tau = 10$ days).
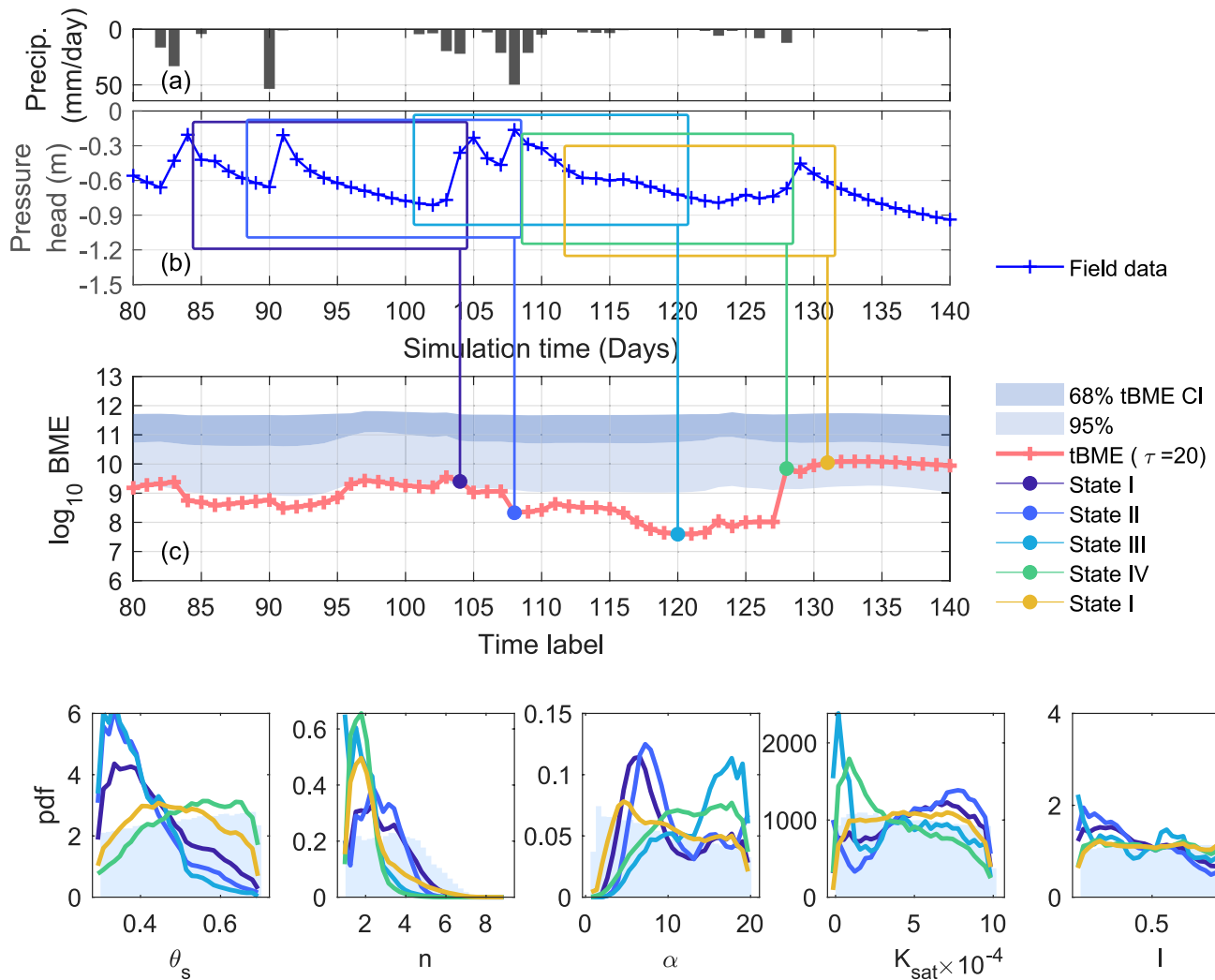
Figure 24. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 20$ days (real observation data).

5. Summary and Conclusions

Model errors are omnipresent in hydrological modeling. Identifying such errors is vital to deepen our system understanding, to appropriately judge the skill of hydrological models, and to improve their forecast reliability in subsequent research. We propose a data-driven method to detect time-dependent model errors. Our method builds on the statistically rigorous Bayesian framework and does not require any prior assumptions about the type or pattern of model error. The novelty of our method is to apply a sequence of Bayesian hypothesis tests by sliding a calibration time window through the given dataset. By testing the assumption that the model has generated the data along the time axis, we can identify the time points when the model fails to describe the system sufficiently well. The core idea of our approach is hence to perform a time-windowed Bayesian analysis that identifies short periods in time where the model statistically disqualifies itself as (pseudo-) true.

We track model performance, measured as Bayesian model evidence (BME), over time to detect the onset of model error, as well as potential error compensation during the error period. As a threshold for reliable error detection, we introduce reference intervals for time-windowed BME (tBME) by sampling the model's predictive distribution as modeled observation data. If a tBME value falls into this reference range, the model's behavior in this time window is interpreted to be quasi-true; if the tBME value falls outside of that range, statistically-significant model error is detected. If tBME drops from the high-density region of the sampling distribution to a very low quantile, this can also be suspected a sign of model error; statistically, this does not reject the hypothesis that

the model structure has generated the data, but it rejects a large part of the model's prior parameter space for that time window.

Our proposed approach is designed to detect error occurrence on various temporal scales, which is especially useful in hydrological modeling. As a side product, we obtain a time sequence of posterior parameter distributions (or of best-fit calibration parameters) that help investigate the reasons for model error and provide guidance for model improvement. Inspecting the dynamic parameter distributions (and, in the specific soil hydraulic context: posterior soil hydraulic functions) allows to put the detected model error into context, and the sensitivity of individual parameters to the sliding calibration time window can provide information about the type and source of error.

We have first provided guidance on how to interpret resulting tBME curves theoretically: their expected behavior is determined by the interplay of the chosen time-window length τ, the residual period length (duration of significant residuals between model and data) L_e, and the residual recurrence interval R. The proposed method works best under the following conditions:

1. The data set should cover the full residual period to be detected.
2. The signal-to-noise ratio in residuals needs to be reasonably high (and ideally error and noise show on different time scales).
3. If potential parameter compensation effects shall be detected, the sliding time-window needs to be shorter than the residual period.
4. If multiple errors shall be identified individually, the recurrence interval of residuals should be larger than the time-window size (although superimposed errors can also be detected, probably with varying degree of clarity in practical applications).

Since the tBME curve purely reflects the residuals, its interpretation regarding potential error sources requires expert knowledge due to potential delays of the system response (e.g., time lags due to storage effects).

Then, we have "calibrated" the reader on how to interpret tBME curves under known error scenarios by designing three synthetic test cases that vary in the error source and error time scales, including a scenario of superimposed errors. Since superposition of two errors of the same sign (both overestimating or underestimating the target value) naturally increases the net residuals (and simplifies the task for our proposed model error indicator), we here deliberately picked error types that could potentially cancel out.

The test cases are inspired by a real case study on soil hydraulic modeling of the Spydia experiment site in the northern Lake Taupo catchment, New Zealand. We have used daily measurements of tensiometric pressure head of the top soil layer to detect (a) hypothetical structural error (single-porosity model vs. dual-porosity system behavior), (b) forcing error (model input not representative of actual precipitation on-site), and (c) a superposition of both error types. Finally, we have presented insights from applying our proposed method to the actual field observations.

Results of our synthetic test cases demonstrate that we can detect the start and end of residual periods for all investigated error sources and temporal scales. Even for superimposed errors that tend to cancel out, we are able to reveal the "fingerprint" of each error with our method. With the sampling distribution of tBME as a reference, we can clearly differentiate between true error signals and model-internal variations that lead to fluctuations in the tBME curve. Varying the time-window size τ has been found to be very insightful, as it intensifies true error signals and filters out noise.

The theoretically expected patterns and the timing of the tBME curve could be discovered from all investigated test cases. Of course, our theoretical considerations hold for idealized, constant residuals during the error period; in reality, we have arbitrary patterns of residuals that affect the shape of the tBME curve. In hydrology in specific, we tend to have rising and falling residuals due to the memory in the system and the model. This means that the timing of, for example, the lowest point of tBME might be different in real cases. However, we found that the expected patterns are surprisingly clear to identify even from real data. Hence, we can use the simple calculation $L_e = L_s - \tau$ to determine the temporal scale of the detected error, that is, the length of the residual period L_e (producing significant residuals between model predictions and observed data) is equal to the total tBME signal length L_s minus the chosen length of the moving time window τ.

In application to the real dataset, our analysis pointed to the model's weaknesses with respect to soil moisture recession and revealed very short temporal scales of error. Superimposed errors of longer scale were not identified; overall, the hypothesis that this model has generated the data could not be rejected during most periods. This, in turn, means that the model is doing respectably well most of the time under the assumed scenario of measurement error and parameter uncertainty. For real-world predictions, we might want to reduce both sources of uncertainty to obtain more precise predictions; that, however, could possibly lead to statistical rejection of the model during larger time periods, because then the hypothesis test through the tBME analysis becomes more strict. Such investigations are left for future studies.

In principle, our concept of tBME could be generalized to a time-windowed calibration analysis using any other model performance metric based on the residuals between model predictions and data. However, we prefer using BME as a model error indicator because it is consistent with the surrounding Bayesian framework, it takes into account measurement noise, and it is cheap to obtain upon calculation of likelihoods that are needed for Bayesian calibration (i.e., to obtain posterior parameter and predictive distributions).

Overall, we have shown the theoretical rigor of the proposed method, and its usefulness for detecting model errors in various scenarios. We are confident that the concept of tBME analysis will be valuable in a wide range of disciplines that rely on predictions by dynamic models with potential model errors.

Appendix A: Supporting Information: Further Results

A1. Synthetic Case 1, $\tau = 10$ Days

Figure A1 presents the 95% posterior credible intervals of the water retention curves and the unsaturated hydraulic conductitvity functiFions at state I and state III* for a window size of $\tau = 10$ days. Comparing the uncertainty intervals of these two states in Figure A1a, we observe slightly decreasing uncertainty when the time window moves from the error-free state (state I) to a parameter compensation state (state III*). This is caused by the observation points leaving the high-density area of the predictive prior.
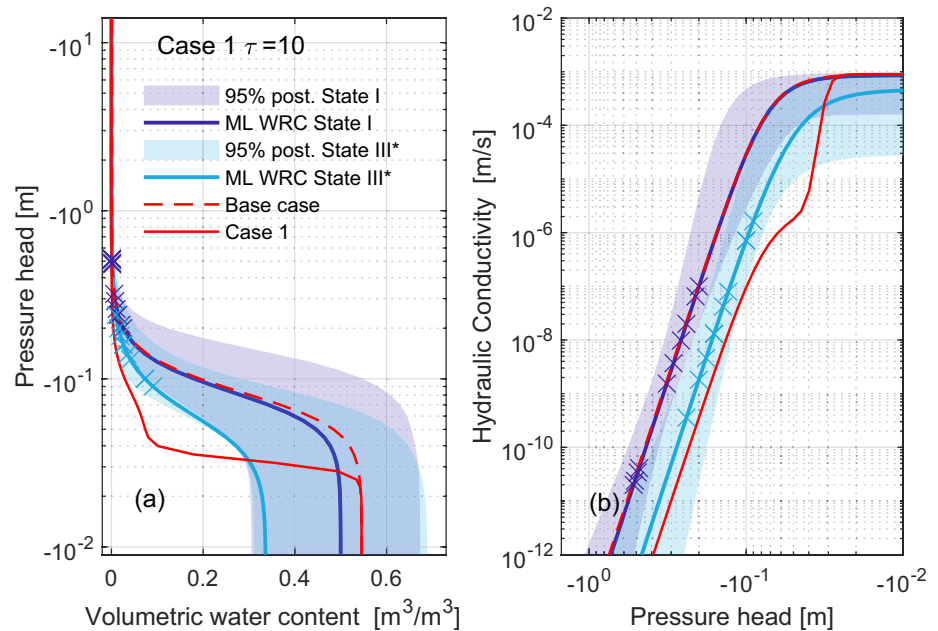

Figure A1. Comparison of the uncertainty range of WRC at state I and state III* in Figure 8 (synthetic case 1 $\tau = 10$).
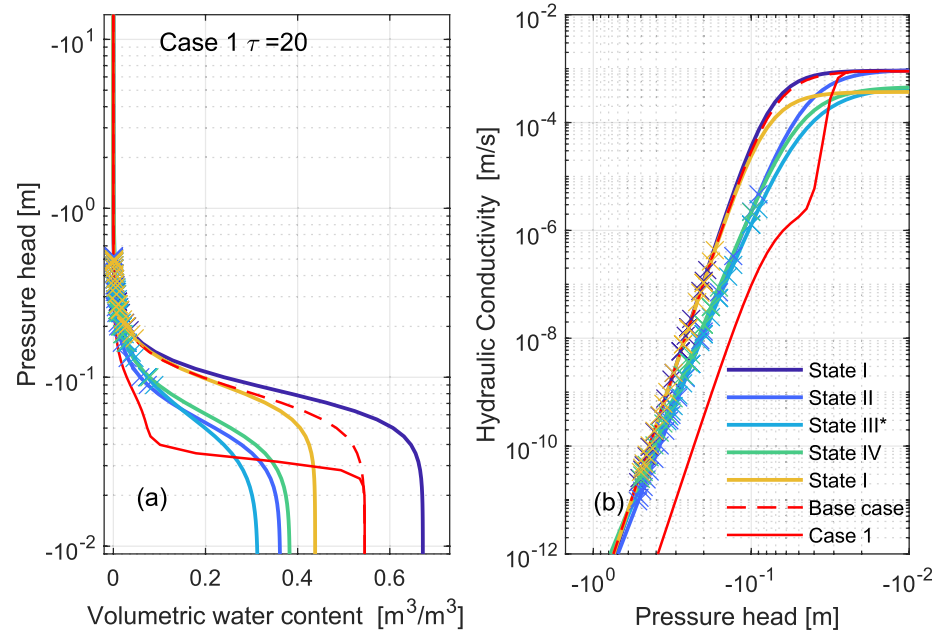
A2. Synthetic Case 1, $\tau = 20$ Days

Figure A2. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 8b (synthetic case 1, window size $\tau = 20$ days). Dashed red lines show true curves of the base case; solid red lines show the curves corresponding to the perturbed dataset.

A3. Synthetic Case 2, $\tau = 20$

We also note the familiar pattern of two drops in the tBME curve (e.g., at $t = 100$ days for $\tau = 10$ days) flattening out to be one single valley for $\tau = 20$ days. This pattern indicates that the time span of model error is between those window sizes. In other words, only states II (initial decline) and IV (final rise) are clearly observed now, but state III (intermediate rise and fall) is absent. This can be investigated in more detail by comparing the zoomed-in tBME curves for $\tau = 10$ and $\tau = 20$ days (Figure 12 and Figure A4, respectively) during days 90–110. The tBME curve for $\tau = 10$ shows 2 drops (state III* identifiable), while only one wide valley appears in the tBME curve for $\tau = 20$ (state III). This signal merge occurs when the time window is larger than the error period itself, because compensation of model error through parameter adjustment cannot happen, and so the tBME curve does not return to the high-density region of the tBME sampling distribution. Hence, it can be judged from Figures 11f and A4c that the error persists for at least 10 and less than 20 days. This conclusion is correct - we can verify it from estimating the length of significant residuals (gray bars in Figure 11b) produced by the second error period which is 13 days. This number is also obtained when determining $L_e = L_s - \tau$ based on theoretical considerations, with $L_s = 33$. Figure A5 shows the maximum-likelihood water retention curve and unsaturated hydraulic conductivity functions of the top soil layer for the five corresponding states. Figure A6 shows the posterior credible intervals of water retention curves and unsaturated hydraulic conductivity curves of the top soil layer for states I and III*.

Figure A7 compares the uncertainty range of WRC at state I and state III* in Figure 16 (synthetic case 3 $\tau = 10$). Figure A8 compares the uncertainty range of WRC at state I and state III in Figure 17 (synthetic case 3 $\tau = 20$).
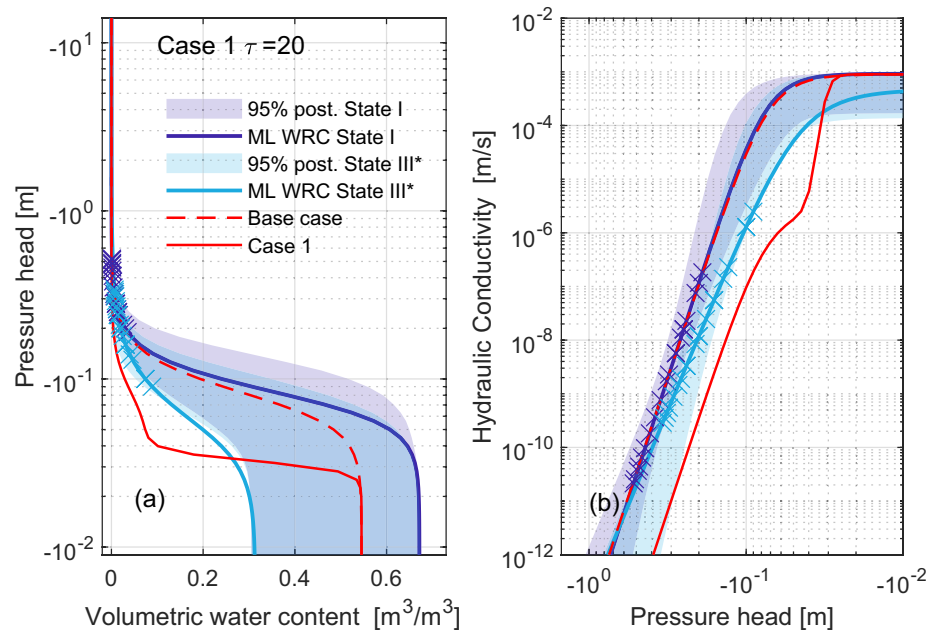
Figure A3. Posterior credible intervals of water retention curves (left) and unsaturated hydraulic conductivity curves (right) of the top soil layer for states I and III* (synthetic case 1, window size $\tau = 20$ days).

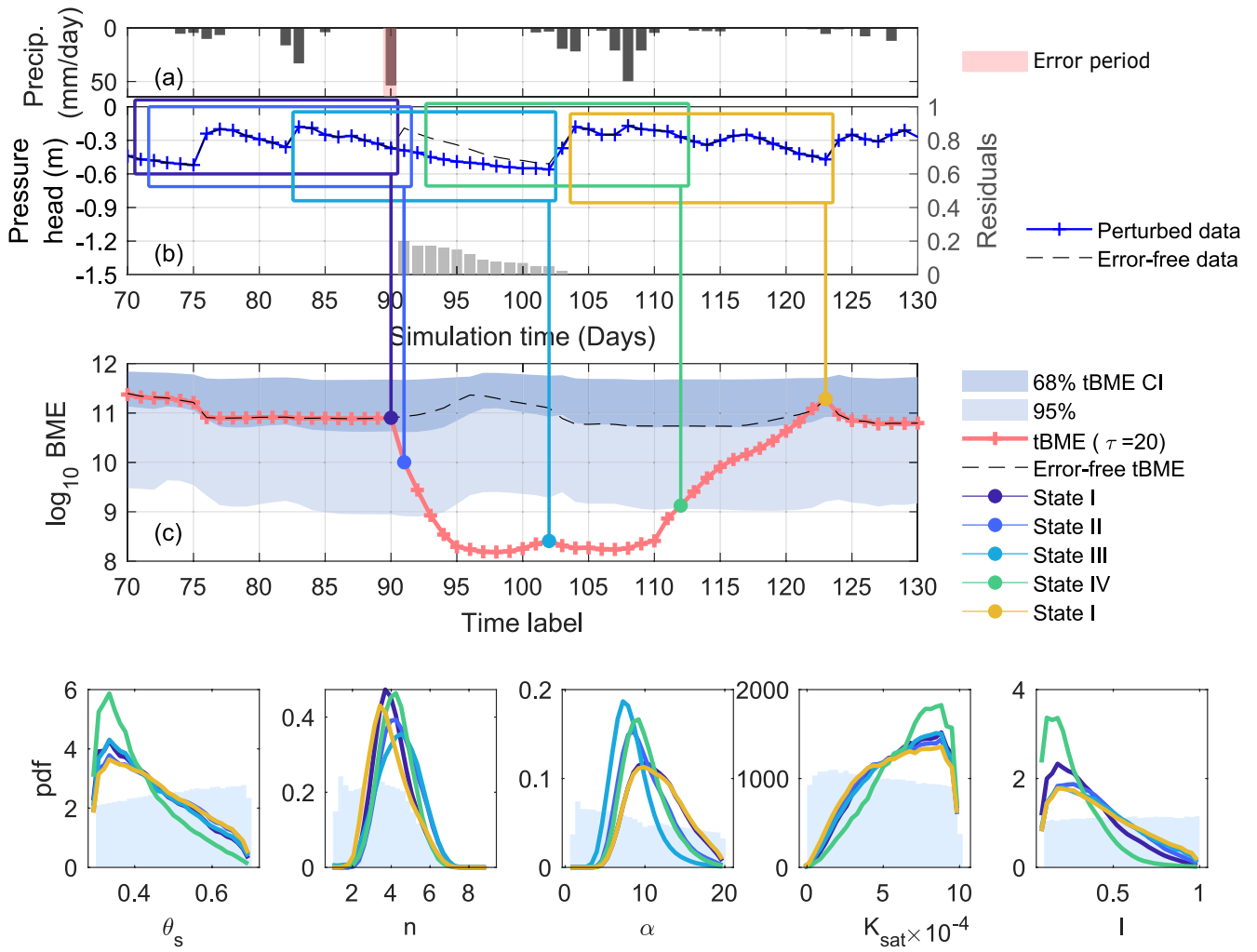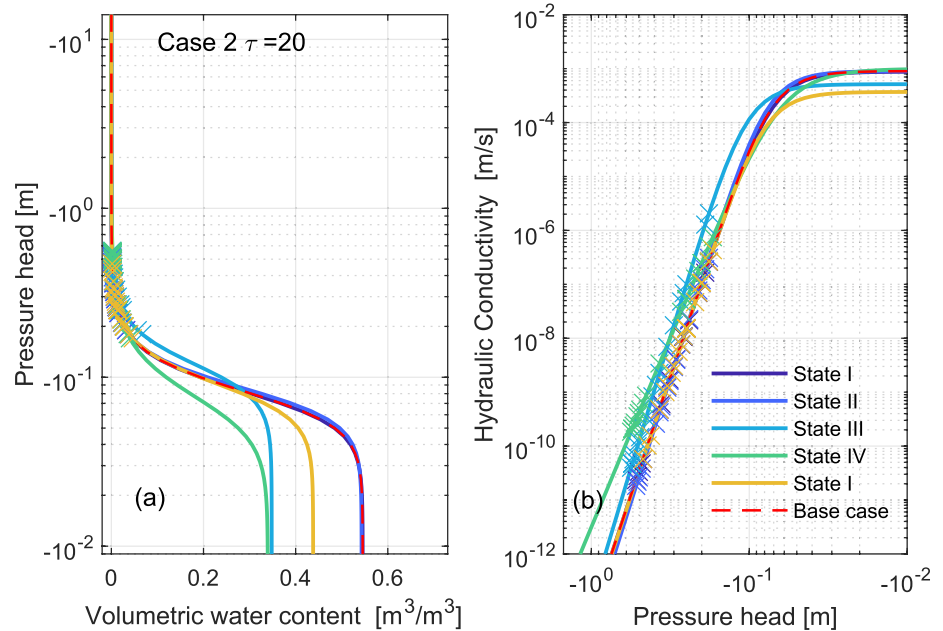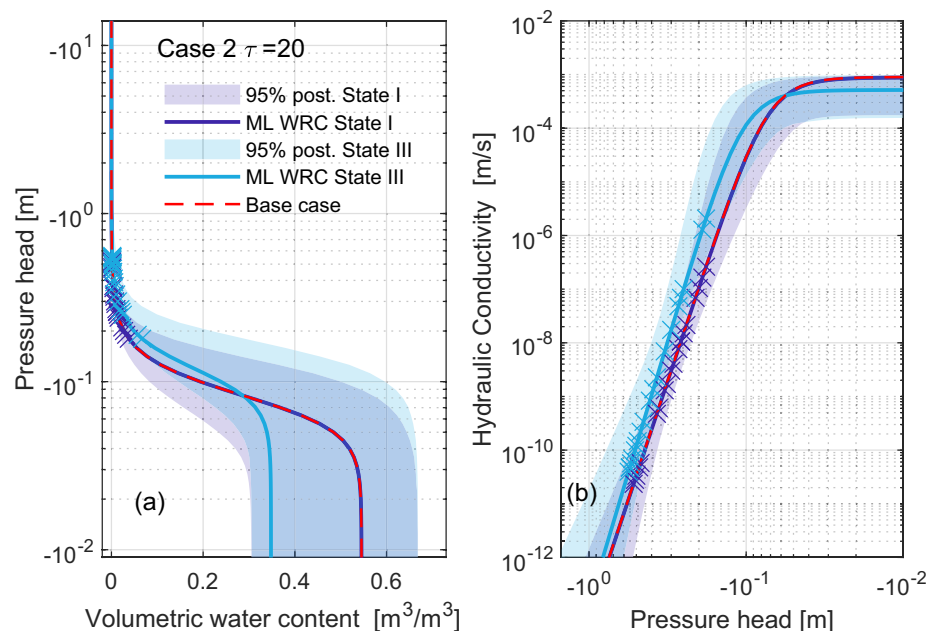### A4. Synthetic Case 3, $\tau = 10$

Figure A4. Zoomed-in view of tBME curve and dynamic posterior parameter distributions for window size $\tau = 20$ days (synthetic case 2).

A5. Synthetic Case 3, $\tau = 20$

Figure A5. Maximum-likelihood water retention curves (left) and unsaturated hydraulic conductivity functions (right) of the top soil layer for the five states color-coded in Figure 12b (synthetic case 2, window size $\tau = 20$ days). Dashed red lines show true curves of the base case; there are no "true" curves for the perturbed dataset because forcing error does not impact the soil hydraulic characteristics.

A6. Real Data, $\tau = 10$

Figure A9, presents the 99% uncertainty range of WRC at state I and state III in Figure 21 which is the real data case analyzed with $\tau = 10$. Figure A11 presents the comparison of the uncertainty range of WRC at state I and state III in Figure 24 (real data, $\tau = 20$). Figure A12 compares the 99% uncertainty range of WRC at state I and state III in Figure 24 (real data, $\tau = 20$).

Figure A6. Posterior credible intervals of water retention curves (left) and unsaturated hydraulic conductivity curves (right) of the top soil layer for states I and III* (synthetic case 2, window size $\tau = 20$ days).
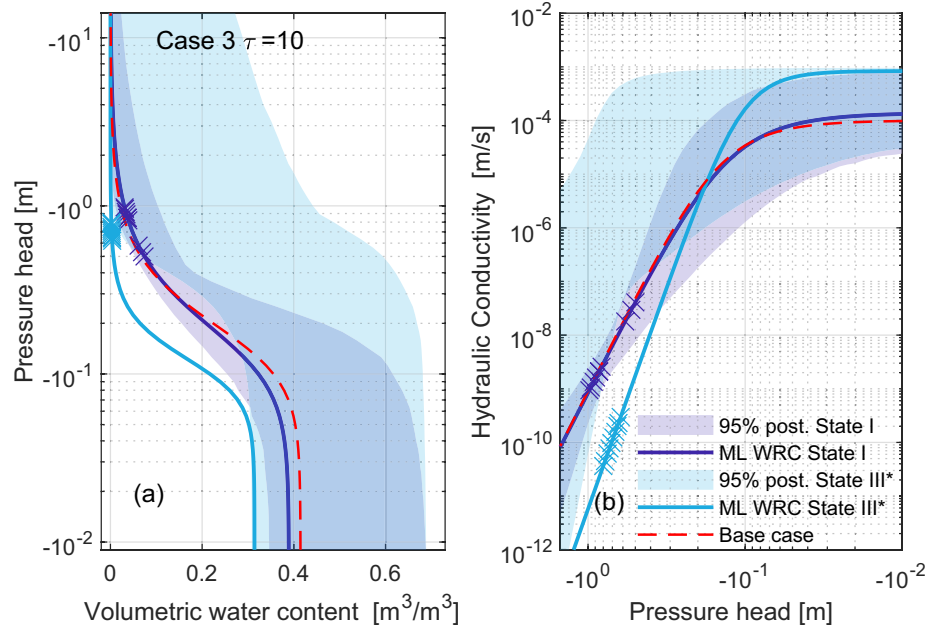
A7. Real Data, $\tau = 20$

Figure A7. Comparison of the uncertainty range of WRC at state I and state III* in Figure 16 (synthetic case 3 $\tau = 10$).
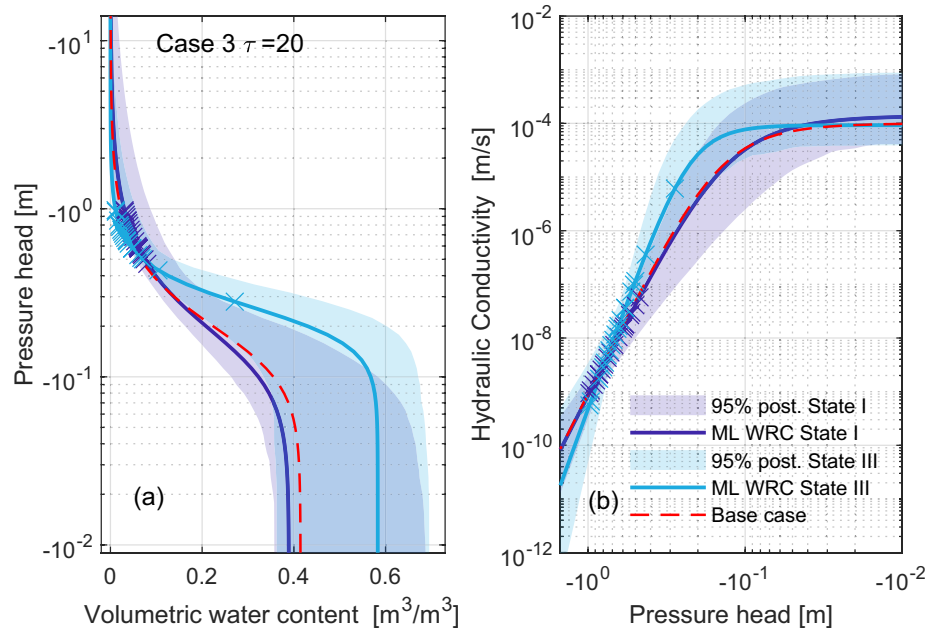

Figure A8. Comparison of the uncertainty range of WRC at state I and state III in Figure 17 (synthetic case 3 $\tau = 20$).
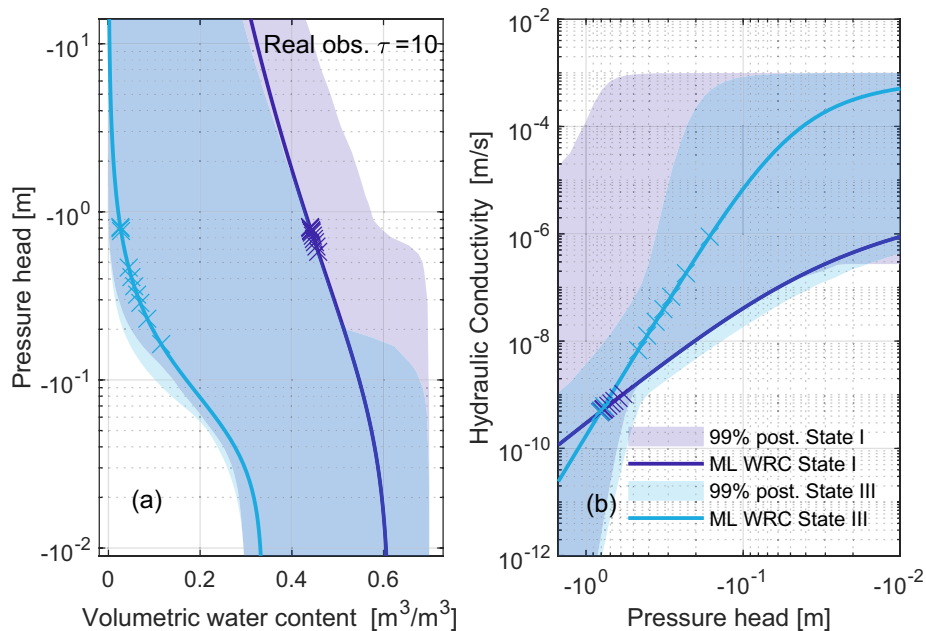
Figure A9. Comparison of the 99% uncertainty range of WRC at state I and state III in Figure 21 (Real data, $\tau = 10$).
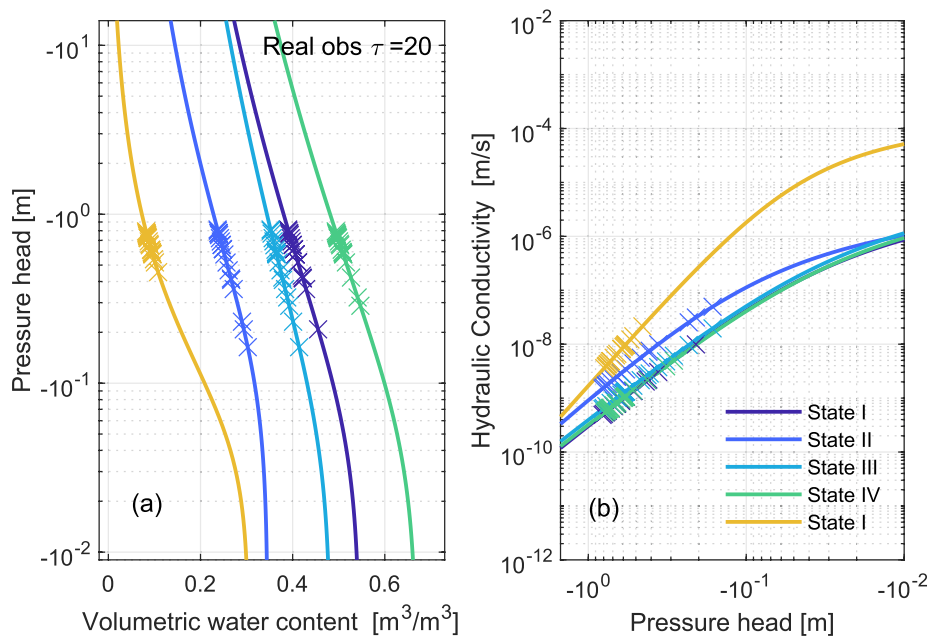

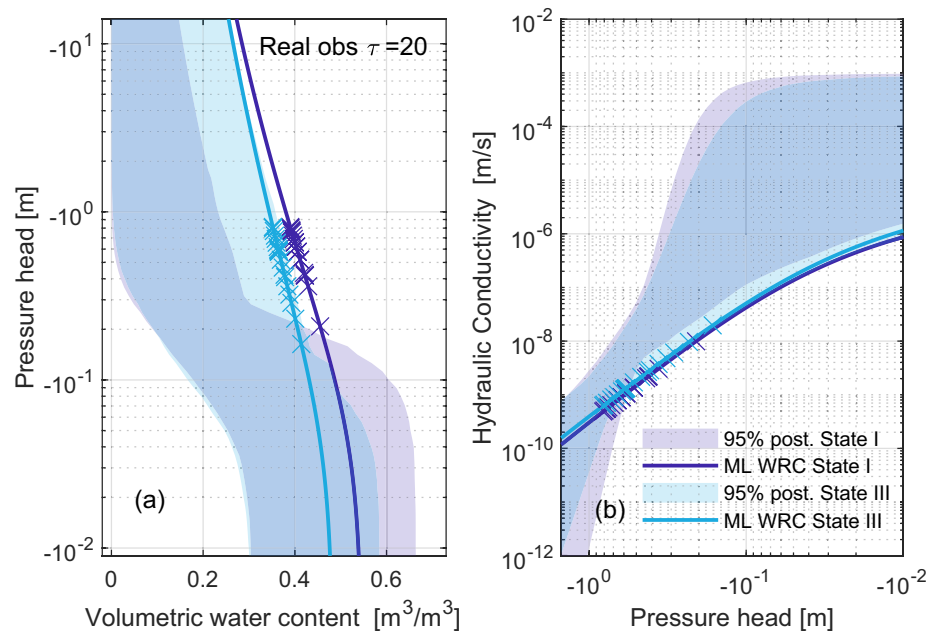**Figure A10.** Comparison of the WRC of MAP at each state in Figure 24 (real data, $\tau = 20$).

Figure A11. Comparison of the uncertainty range of WRC at state I and state III in Figure 24 (real data, $\tau = 20$).
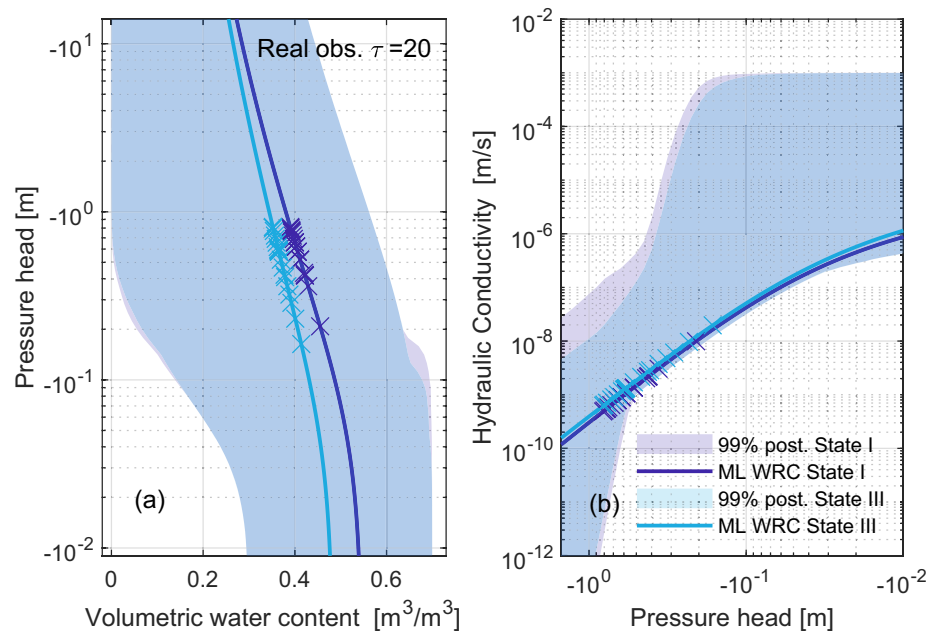

Figure A12. Comparison of the 99% uncertainty range of WRC at state I and state III in Figure 24 (real data, $\tau = 20$).

Data Availability Statement

All code files of the analysis in this research are provided on the platform of the data repository of the University of Stuttgart (DaRUS) https://doi.org/10.18419/darus-1836.

References

Ajami, N. K., Duan, Q., & 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). https://doi.org/10.1029/2005WR004745

Ammann, L., Fenicia, F., & Reichert, P. (2019). A likelihood framework for deterministic hydrological models and the importance of non-stationary autocorrelation. *Hydrology and Earth System Sciences*, *23*(4), 2147–2172. https://doi.org/10.5194/hess-23-2147-2019

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., et al. (2009). Hess opinions crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences*, *13*(10), 1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Barkle, G., Wöhling, T., Stenger, R., Mertens, J., Moorhead, B., Wall, A., & Clague, J. (2011). Automated equilibrium tension lysimeters for measuring water fluxes through a layered, volcanic vadose profile in New Zealand. *Vadose Zone Journal*, *10*(2), 747–759. https://doi.org/10.2136/vzj2010.0091

Bastola, S., Ishidaira, H., & Takeuchi, K. (2008). Regionalisation of hydrological model parameters under parameter uncertainty: A case study involving top model and basins across the globe. *Journal of Hydrology*, *357*(3), 188–206. https://doi.org/10.1016/j.jhydrol.2008.05.007

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. *WIREs Water*, *5*(3), e1278. https://doi.org/10.1002/wat2.1278

Box, G. E., & Tiao, G. C. (1992). *Nature of Bayesian inference*, John Wiley & Sons, Ltd. (pp. 1–75). https://doi.org/10.1002/9781118033197.ch1

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*(11), 114007. https://doi.org/10.1088/0266-5611/30/11/114007

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, *51*(4), 2515–2542. https://doi.org/10.1002/2015WR017200

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006735

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*(5). https://doi.org/10.1029/2011WR011721

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

Del Giudice, D., Reichert, P., Bareš, V., Albert, C., & Rieckermann, J. (2015). Model bias and complexity—Understanding the effects of structural deficits and input errors on runoff predictions. *Environmental Modelling & Software*, *64*, 205–214. https://doi.org/10.1016/j.envsoft.2014.11.006

Demissie, Y. K., Valocchi, A. J., Minsker, B. S., & Bailey, B. A. (2009). Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of Hydrology*, *364*(3–4), 257–271. https://doi.org/10.1016/j.jhydrol.2008.11.007

Dobler, C., Hagemann, S., Wilby, R. L., & Stötter, J. (2012). Quantifying different sources of uncertainty in hydrological projections in an alpine watershed. *Hydrology and Earth System Sciences*, *16*(11), 4343–4360. https://doi.org/10.5194/hess-16-4343-2012

Durner, W. (1994). Hydraulic conductivity estimation for soils with heterogeneous pore structure. *Water Resources Research*, *30*(2), 211–223. https://doi.org/10.1029/93WR02676

Evin, G., Kavetski, D., Thyer, M., & Kuczera, G. (2013). Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resources Research*, *49*(7), 4518–4524. https://doi.org/10.1002/wrcr.20284

Fenicia, F., Kavetski, D., Reichert, P., & Albert, C. (2018). Signature-domain calibration of hydrological models using approximate Bayesian computation: Empirical analysis of fundamental properties. *Water Resources Research*, *54*(6), 3958–3987. https://doi.org/10.1002/2017wr021616

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*(11). https://doi.org/10.1029/2010WR010174

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press. https://doi.org/10.1201/b16018

Getahun, Y. S., & Van Lanen, H. (2015). Assessing the impacts of land use-cover change on hydrology of Melka Kuntrie subbasin in Ethiopia, using a conceptual hydrological model. *Hydrology: Current Research*, *6*(3), 1. https://doi.org/10.4172/2157-7587.1000210

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8). https://doi.org/10.1029/2011WR011044

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

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

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*(3), 1688–1715. https://doi.org/10.1002/2017wr021902

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., et al. (2014). Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resources Research*, *50*(9), 7445–7469. https://doi.org/10.1002/2014wr015484

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005wr004376

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004362

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., & Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*, *56*(9), e2019WR025975. https://doi.org/10.1029/2019WR025975

Kuczera, G., Kavetski, D., Franks, S., & Thyer, M. (2006). Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology*, *331*(1), 161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010

Lee, S., Yen, H., Yeo, I.-Y., Moglen, G. E., Rabenhorst, M. C., & McCarty, G. W. (2020). Use of multiple modules and Bayesian model averaging to assess structural uncertainty of catchment-scale wetland modeling in a coastal plain landscape. *Journal of Hydrology*, *582*, 124544. https://doi.org/10.1016/j.jhydrol.2020.124544

Leube, P., Geiges, A., & Nowak, W. (2012). Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. *Water Resources Research*, *48*(2). https://doi.org/10.1029/2010WR010137

Liu, J. S. (2004). *Monte Carlo strategies in scientific computing*. Springer. https://doi.org/10.1007/978-0-387-76371-2

Liu, P., Elshall, A. S., Ye, M., Beerli, P., Zeng, X., Lu, D., & Tao, Y. (2016). Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resources Research*, *52*(2), 734–758. https://doi.org/10.1002/2014WR016718

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *WIREs Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Menberg, K., Heo, Y., & Choudhary, R. (2019). Influence of error terms in Bayesian calibration of energy system models. *Journal of Building Performance Simulation*, *12*(1), 82–96. https://doi.org/10.1080/19401493.2018.1475506

Mirzaei, M., Huang, Y. F., El-Shafie, A., & Shatirah, A. (2015). Application of the generalized likelihood uncertainty estimation (glue) approach for assessing uncertainty in hydrological models: A review. *Stochastic Environmental Research and Risk Assessment*, *29*(5), 1265–1273. https://doi.org/10.1007/s00477-014-1000-6

Monteith, J. L. (1981). Evaporation and surface temperature. *Quarterly Journal of the Royal Meteorological Society*, *107*(451), 1–27. https://doi.org/10.1002/qj.49710745102

Morán-Tejeda, E., Zabalza, J., Rahman, K., Gago-Silva, A., López-Moreno, J. I., Vicente-Serrano, S., et al. (2015). Hydrological impacts of climate and land-use changes in a mountain watershed: Uncertainty estimation based on model comparison. *Ecohydrology*, *8*(8), 1396–1416. https://doi.org/10.1002/eco.1590

Motavita, D., Chow, R., Guthke, A., & Nowak, W. (2019). The comprehensive differential split-sample test: A stress-test for hydrological model robustness under climate variability. *Journal of Hydrology*, *573*, 501–515. https://doi.org/10.1016/j.jhydrol.2019.03.054

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, *12*(3), 513–522. https://doi.org/10.1029/WR012i003p00513

Mustafa, S. M. T., Nossent, J., Ghysels, G., & Huysmans, M. (2018). Estimation and impact assessment of input and parameter uncertainty in predicting groundwater flow with a fully distributed model. *Water Resources Research*, *54*(9), 6585–6608. https://doi.org/10.1029/2017WR021857

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for earth science? Hypothesis testing. *Water Resources Research*, *56*(2), e2019WR024918. https://doi.org/10.1029/2019WR024918

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*(9), 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

Neuweiler, I., & Helmig, R. (2017). Debates—Hypothesis testing in hydrology: A subsurface perspective. *Water Resources Research*, *53*(3), 1784–1791. https://doi.org/10.1002/2016wr020047

Nowak, W., & Guthke, A. (2016). Entropy-based experimental design for optimal model discrimination in the geosciences. *Entropy*, *18*(11), 409. https://doi.org/10.3390/e18110409

Oladyshkin, S., & Nowak, W. (2019). The connection between Bayesian inference and information theory for model selection, information gain and experimental design. *Entropy*, *21*(11), 1081. https://doi.org/10.3390/e21111081

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Pathiraja, S., Marshall, L., Sharma, A., & Moradkhani, H. (2016). Hydrologic modeling in dynamic catchments: A data assimilation approach. *Water Resources Research*, *52*(5), 3350–3372. https://doi.org/10.1002/2015wr017192

Pianosi, F., & Wagener, T. (2016). Understanding the time-varying importance of different uncertainty sources in hydrological modelling using global sensitivity analysis. *Hydrological Processes*, *30*(22), 3991–4003. https://doi.org/10.1002/hyp.10968

Ragab, R., Kaelin, A., Afzal, M., & Panagea, I. (2020). Application of generalized likelihood uncertainty estimation (glue) at different temporal scales to reduce the uncertainty level in modelled river flows. *Hydrological Sciences Journal*, *65*(11), 1856–1871. https://doi.org/10.1080/02626667.2020.1764961

Reichert, P., & Mieleitner, J. (2009). Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Research*, *45*(10). https://doi.org/10.1029/2009WR007814

Reichert, P., & Schuwirth, N. (2012). Linking statistical bias description to multiobjective model calibration. *Water Resources Research*, *48*(9). https://doi.org/10.1029/2011WR011391

Remondi, F., Kirchner, J. W., Burlando, P., & Fatichi, S. (2018). Water flux tracking with a distributed hydrological model to quantify controls on the spatio-temporal variability of transit time distributions. *Water Resources Research*, *54*(4), 3081–3099. https://doi.org/10.1002/2017WR021689

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., & Franks, S. W. (2010). Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2009WR008328

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information theory for model diagnostics: Structural error is indicated by trade-off between functional and predictive performance. *Water Resources Research*, *55*(8), 6534–6554. https://doi.org/10.1029/2018wr023692

Sargsyan, K., Huan, X., & Najm, H. N. (2019). Embedded model error representation for Bayesian model calibration. *International Journal for Uncertainty Quantification*, *9*(4), 365–394. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2019027384

Schöniger, A., Wöhling, T., & Nowak, W. (2015). A statistical concept to assess the uncertainty in Bayesian model weights and its impact on model ranking. *Water Resources Research*, *51*(9), 7524–7546. https://doi.org/10.1002/2015WR016918

Schöniger, A., Wöhling, T., Samaniego, L., & 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. https://doi.org/10.1002/2014wr016062

Simunek, J., Van Genuchten, M. T., & Sejna, M. (2005). *The HYDRUS-1D software package for simulating the one-dimensional movement of water, heat, and multiple solutes in variably-saturated media*, University of California-Riverside Research Reports. (Vol. 3, pp. 1–240).

Smith, T., Marshall, L., & Sharma, A. (2015). Modeling residual hydrologic errors with Bayesian inference. *Journal of Hydrology*, *528*, 29–37. https://doi.org/10.1016/j.jhydrol.2015.05.051

Steinschneider, S., Wi, S., & Brown, C. (2015). The integrated effects of climate and hydrologic uncertainty on future flood risk assessments. *Hydrological Processes*, *29*(12), 2823–2839. https://doi.org/10.1002/hyp.10409

Teweldebrhan, A. T., Burkhart, J. F., & Schuler, T. V. (2018). Parameter uncertainty analysis for an operational hydrological model using residual-based and limits of acceptability approaches. *Hydrology and Earth System Sciences*, *22*(9), 5021–5039. https://doi.org/10.5194/hess-22-5021-2018

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, *44*(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Vogel, T., & Cislerova, M. (1988). On the reliability of unsaturated hydraulic conductivity calculated from the moisture retention curve. *Transport in Porous Media*, *3*(1), 1–15. https://doi.org/10.1007/BF00222683

Wagner, P. D., Bhallamudi, S. M., Narasimhan, B., Kantakumar, L. N., Sudheer, K., Kumar, S., et al. (2016). Dynamic integration of land use changes in a hydrologic assessment of a rapidly developing Indian catchment. *The Science of the Total Environment*, *539*, 153–164. https://doi.org/10.1016/j.scitotenv.2015.08.148

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S. (2001). A framework for development and application of hydrological models. *Hydrology and Earth System Sciences*, *5*(1), 13–26. https://doi.org/10.5194/hess-5-13-2001

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., & Gupta, H. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, *17*(2), 455–476. https://doi.org/10.1002/hyp.1135

Westra, S., Thyer, M., Leonard, M., Kavetski, D., & Lambert, M. (2014). A strategy for diagnosing and interpreting hydrological model nonstationarity. *Water Resources Research*, *50*(6), 5090–5113. https://doi.org/10.1002/2013wr014719

Wilby, R. L. (2005). Uncertainty in water resource model parameters used for climate change impact assessment. *Hydrological Processes*, *19*(16), 3201–3219. https://doi.org/10.1002/hyp.5819

Wöhling, T., Schöniger, A., Gayler, S., & 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. https://doi.org/10.1002/2014WR016292

Wöhling, T., Schütze, N., Heinrich, B., imnek, J., & Barkle, G. F. (2009). Three-dimensional modeling of multiple automated equilibrium tension lysimeters to measure vadose zone fluxes. *Vadose Zone Journal*, *8*(4), 1051–1063. https://doi.org/10.2136/vzj2009.0040

Wöhling, T., & Vrugt, J. A. (2008). Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2008WR007154

Wöhling, T., Vrugt, J. A., & Barkle, G. F. (2008). Comparison of three multiobjective optimization algorithms for inverse modeling of vadose zone hydraulic properties. *Soil Science Society of America Journal*, *72*(2), 305–319. https://doi.org/10.2136/sssaj2007.0176

Xu, T., & Valocchi, A. J. (2015). A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, *51*(11), 9290–9311. https://doi.org/10.1002/2015WR017912

Xu, T., Valocchi, A. J., Ye, M., & Liang, F. (2017). Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. *Water Resources Research*, *53*(5), 4084–4105. https://doi.org/10.1002/2016wr019831