Reviews of Geophysics / Volume 57, Issue 1 / p. 146-182

Review Article

Free Access

Beyond Classical Observations in Hydrogeology: The Advantages of Including Exchange Flux, Temperature, Tracer Concentration, Residence Time, and Soil Moisture Observations in Groundwater Model Calibration

Oliver S. Schilling, Peter G. Cook, Philip Brunner

First published: 20 February 2019
https://doi.org/10.1029/2018RG000619
Citations: 12

Abstract

Traditionally, groundwater and surface water flow models have been calibrated against two observation types: hydraulic heads and surface water discharge. It has repeatedly been demonstrated, however, that these classical observations do not contain sufficient information to calibrate flow models. To reduce the predictive uncertainty of flow models, the consideration of other observation types constitutes a promising way forward. Despite the ever-increasing availability of other observation types, however, they are still unconventional when it comes to flow model calibration. By reviewing studies that included nonclassical observations in flow model calibration, benefits and challenges associated with their integration in flow model calibration were identified, and their information content was analyzed. While explicit simulation of mass transport processes in flow models poses challenges, even simplified approaches to integrate tracer concentrations yield significantly better calibration results than using only classical observations. For a majority of calibrated flow models, observations of tracer concentrations and of exchange fluxes were beneficial. Temperature observations improved the simulation of heat transport but often worsened all other model outcomes. Only when temperature observations were made within 2 m of the surface water-groundwater interface did they have the potential to also improve flow and mass transport simulations. Surprisingly, many models were calibrated manually rather than with the widely available, mathematically robust and automated tools. There is a clear need for more systematic implementation of unconventional observations and automated flow model calibration as well as for more systematic quantification of the information content of unconventional observations.

Plain Language Summary

Traditionally, groundwater and surface water flow models, which are critical for water resources assessment, have been calibrated against only two classical observation types: groundwater levels and surface water discharge. In the past, it has repeatedly been demonstrated that these classical observations do not contain sufficient information to calibrate the parameters required for the simulation of groundwater and surface water flow systems. Owing to the rapid development of measurement techniques throughout the last three decades, however, many other observations of hydrological systems have become widely available. Despite this, observation types other than the classical ones are still unconventional when it comes to flow model calibration. The overall goal of this review is to identify optimal observation types and procedures for flow model calibration and hydrological predictions. We found that observations of tracer concentrations and exchange fluxes are beneficial for most flow models. Temperatures improve the simulation of heat transport but often worsen other flow model outcomes, unless temperatures are measured within 2 m of the surface water-groundwater interface. We identified a need for more systematic implementation of unconventional observations in flow model calibration and for more systematic quantification of the information content of unconventional observations.

1 Introduction

Groundwater (GW) flow models are an important tool for the characterization and management of GW resources (Anderson et al., 2015; Freeze et al., 1990; Jakeman et al., 2016; Poeter & Hill, 1997). In process-based or physically-based GW flow models, the governing equation to simulate flow of water in the subsurface is Darcy's law, which is typically solved numerically on a finite-differences or finite-element grid. GW flow models are used to simulate the movement of water in the subsurface both in steady state and transiently and range in scale from small soil columns to large watershed models. GW flow models require a significant number of parameters to be defined, but due to the complex and inaccessible nature of the subsurface, these parameters are difficult to measure and therefore rarely known (Anderson et al., 2015; Poeter & Hill, 1997). Due to the limited ability to measure the true parameter values, especially of strongly heterogeneous properties such as the hydraulic conductivity (K) or the effective porosity (f), GW flow models have to be calibrated against observations of system states (Anderson et al., 2015; Hrachowitz & Clark, 2017). In theory, if the set of observations that is used to calibrate a flow model contained sufficient information to constrain all unknown parameters, the inverse problem would be well posed, and the underlying parameter values could be uniquely identified (Aster et al., 2013). However, due to the large complexity and spatial and temporal dynamics of natural hydrological and hydrogeological systems, the observations of system states that we are able to make never contain sufficient information to uniquely identify all unknown system properties. Inevitably, the results of flow model calibration are nonunique parameter sets that are associated with uncertainty (Anderson et al., 2015; Beven, 2006; de Marsily et al., 2000; Doherty, 2015; Hill & Tiedeman, 2007; McLaughlin & Townley, 1996; Moore et al., 2010; Moore & Doherty, 2006). While zonation of the subsurface to homogeneous units creates parameter uniqueness through a reduction of the degrees of freedom during flow model calibration, this practice does not change our limited knowledge about the subsurface and, therefore, does not eliminate the uncertainty of predictions made with flow models (Anderson et al., 2015; Doherty, 2015). This is true for all forms of artificial parameter reduction, that is, for all forms of regularization (Doherty, 2015; Moore et al., 2010; Moore &

Doherty, 2006).

The problem of parameter nonuniqueness may be less acute for GW flow systems with close to homogeneous properties or systems, which are not dominated by geological heterogeneity or complex hydrogeological processes. But with the capabilities of the current generation of flow models, one is no longer restricted to the simulation of such simple GW systems anymore or forced to oversimplify more complex GW systems—the current trend is toward the simulation of complex systems based on integrated surface water (SW)-GW flow models (IFMs; Barthel & Banzhaf, 2016; Paniconi & Putti, 2015). Compared to numerical models that exclusively simulate GW flow, IFMs enable the simulation of GW and SW flow in a physically-based and fully-coupled way and allow the inclusion of many hydrologically relevant processes such as unsaturated flow through complex heterogeneous structures (e.g., Irvine et al., 2012; Schilling, Irvine, et al., 2017; Tang et al., 2015, 2018), heat and mass transport (e.g., Carniato et al., 2015; Karan et al., 2014; Kurtz et al., 2014; Schilling et al., 2013), snow accumulation, melt and pore water freeze-thaw (e.g., Cochand et al., 2019; Evans & Ge, 2017; Painter et al., 2016; Schilling et al., 2019; Shojae Ghias et al., 2017), and SW-GW-vegetation interactions (e.g., Banks et al., 2011; Maxwell & Condon, 2016; Schomburg et al., 2018). Compared to numerical flow models that exclusively simulate GW flow, IFMs require more parameters and boundary conditions to be defined and calibrated (the minimally required parameters and boundary conditions of different types of GW and SW simulations are listed in Table 1). Increased model complexity, however, requires more observations on the underlying parameters and processes for robust flow model construction and flow model calibration (Anderson et al., 2015; Doherty & Hunt, 2009). Consequently, there exists a trade-off between the benefits of increased flow model complexity and the increased requirements for observations to define the additional flow model parameters, processes, and boundary conditions. Hrachowitz and Clark (2017) recently discussed the benefits and trade-offs of different model types and modeling philosophies. They concluded that the suitable modeling philosophy, flow model, and model complexity are dependent not only on the purpose of and the predictions to be made with the flow model but also on the available information on the underlying flow system.

Table 1. Summary of Required Parameters and Boundary Conditions to be Defined, Separated by Type of Process

Domain	Process	Parameters	Boundary conditions
GW	Hydraulic head propagation	K, S	H_{BC} and/or Q_{BC}
	Advective flow	K, f	H_{BC} and/or Q_{BC}
	Conservative mass transport	K, f, α_{sol}, D_{sol}	H_{BC} and/or Q_{BC}, C_{BC}
	Reactive mass transport	K, f, α_{sol}, D_{sol}, λ_{source}, λ_{sink}	H_{BC} and/or Q_{BC}, C_{BC}
	Heat transport	K, f, α_{heat}, c_{bulk}, ρ_{bulk}, κ_{bulk},	H_{BC} and/or Q_{BC}, T_{BC}
	GW age simulations	K, f, (α_{sol}, D_{sol}, λ_{source}, λ_{sink})	H_{BC} and/or Q_{BC}, (C_{BC})
SW	Advective flow	n, h_d, h_o	H_{BC} and/or Q_{BC}
	Conservative mass transport	n, h_d, h_o, $\alpha_{sol,SW}$, $D_{sol,SW}$	H_{BC} and/or Q_{BC}, C_{BC}
	Reactive mass transport	n, h_d, h_o, $\alpha_{sol,SW}$, $D_{sol,SW}$, $\lambda_{source,SW}$, $\lambda_{sink,SW}$	H_{BC} and/or Q_{BC}, C_{BC}

Heat transport	n, h_d, h_o, $\alpha_{heat,SW}$, c_{bulk}, ρ_{bulk}, κ_{bulk}	H_{BC} and/or Q_{BC}, T_{BC}

Note. GW = groundwater; SW = surface water. The listed parameters and boundary conditions only represent the ones that are minimally required. Many other parameters and boundary conditions exist. For GW age simulations, parameters required for the simplified, purely advective flow as well as for the explicit advective-dispersive flow (in brackets) are provided. Parameters, boundary conditions and their associated dimensions are explained in the notation table.

Parameter nonuniqueness increases both with increasing complexity of the modeled system and with increasing complexity of the employed numerical model. Given the high likelihood of not being able to identify the true underlying parameters, the goal for the calibration of flow models that exclusively simulate GW flow and of IFMs (which for the rest of this manuscript are grouped under the term "flow models") should be the identification of parametrizations that are appropriate for the intended purpose of the model, rather than identifying the true underlying parameter values (Anderson et al., 2015; Doherty, 2015; Haitjema, 2015). As mentioned before, flow model calibration does allow the range of likely parameter values to be constrained, which, in turn, reduces the uncertainty of predictions made with flow models (Anderson et al., 2015; Doherty, 2015). Besides identifying an appropriate parametrization, the important questions that need to be answered through flow model calibration studies thus are as follows (Anderson et al., 2015; Doherty, 2015): (1) What are the uncertainties of the simulations and predictions made with flow models? (2) How much could the uncertainty of simulations and predictions made with a flow model be reduced through calibration? (3) How much is the information content or worth of different observation types for the reduction of predictive uncertainty, and which parameters are informed by what observation type?

Despite a large body of literature devoted to the topic of flow model calibration, a considerable gap exists between the understanding achieved in the last three decades and the current flow model calibration practice: It is well researched that the ***classical*** observations of hydraulic head (H) and SW discharge (Q_{SW}), which are traditionally used for flow model calibration, do not contain sufficient information to find an appropriate parametrization for most modeling purposes (Anderson et al., 2015; Beven, 2006; Delottier et al., 2016; Doherty & Hunt, 2009; Haitjema, 2006, 2015; Hunt et al., 2006; Jakeman & Hornberger, 1993; Moore et al., 2010; Townley, 2012) and that manual, trial-and-error-type calibration of flow models rarely results in mathematically appropriate parameter sets (Anderson et al., 2015; Doherty, 2015; Doherty & Hunt, 2010; Hill & Tiedeman, 2007; McLaughlin & Townley, 1996; Poeter & Hill, 1997; Townley, 2012; Townley & Wilson, 1985). Despite having reached this understanding, Simmons et al. (2012), Brunner et al. (2017), and Partington et al. (2017) found that flow model calibration is still often based on the classical observations of H and Q_{SW} alone and on manual trial-and-error-type calibration. Limiting the calibration of flow models to classical observations is a stark contrast to the ever-increasing availability of other observations of hydrological systems. Cook and Herczeg (2000), Scanlon et al. (2002), Anderson (2005), Healy and Scanlon (2010), Harvey and Gooseff (2015), Maliva (2016), and Brunner et al. (2017), among others, provide vast information on how to obtain observations of

- **temperature**
- **exchange fluxes** (i.e., infiltration of SW, recharge, GW discharge into rivers or as springs, baseflow [SW flow during low-flow periods], and evapotranspiration [ET])
- **tracer concentrations** (e.g., concentrations of solutes, isotopes, and dissolved gases)

- **temporal information** (i.e., residence times [RTs] and travel times [TTs])

As outlined by these authors, such observations can now be obtained more reliably, at a much higher spatial and temporal resolution and at a fraction of the cost compared to a few decades ago. Despite the fact that these observation types are now widely available, in the *general practice* of flow model calibration, these observations are so rarely used that we herein consider and refer to these observations as *unconventional* for flow model calibration. While not yet being part of the general practice, a considerable amount of literature devoted to the topic of flow modeling has already highlighted that unconventional observation types should be used alongside the classical ones in order to reduce both parameter nonuniqueness and predictive uncertainty (e.g., Anderson et al., 2015; Doherty & Hunt, 2010; Hill & Tiedeman, 2007; Hunt et al., 2006; Simmons et al., 2012; Townley, 2012). What is rarely discussed in the literature devoted to this topic, however, is which observation type to use in what context and how to most efficiently implement unconventional observation types into the flow modeling and the calibration process. Doherty and Hunt (2010) and Hill and Tiedeman (2007) provide some advice on how to include different observation types into the multivariate objective functions of automated flow model calibration routines, but a comprehensive guide on how to best integrate unconventional observation types into flow model calibration is still missing. We suspect that the lack in guidance on which observation type to use when and how is the primary reason for the current underrepresentation of unconventional observations in the general practice of flow model calibration. The following challenges, potential pitfalls, and unresolved issues are associated with the use of unconventional observations for flow model calibration:

1. **Challenges associated with the integration of unconventional observations**: In contrast to the classical observation types that are simulated by flow models by default, other observations may relate to processes that are often not by default simulated by flow models (e.g., heat and mass transport or GW RTs). To make unconventional observations useful for flow model calibration, the processes to which they relate must be an integral part of the flow model. But how should unconventional observations ideally be integrated into flow models? Integrating these additional processes into flow models is not straightforward, as additional processes require more complex conceptual models to be defined—and this is often difficult. Three different approaches to address additional processes can be used: (i) an explicit and physically-based simulation of the process, (ii) a simplified simulation of the process, or (iii) a preprocessing/transformation of the unconventional observations to an observation type that does not require additional processes to be simulated. An explicit simulation provides the physically most appropriate representation of additional processes but also results in the highest number of additional parameters and additional initial and boundary conditions that have to be calibrated or measured and, therefore, the most elaborate conceptual model to be defined (see Table 1). This may result in overparametrization, that is, introduce additional uncertainty compared to the information content of the available observations. Moreover, the explicit simulation of additional processes might result in substantially longer run times and harbors potential for numerical instabilities —both can pose serious problems for model calibration and the quantification of the uncertainty of model predictions. Simplified simulation routines, on the other hand, can provide a more efficient alternative of including additional observation types without introducing too many additional parameters but might result in an oversimplified conceptual model of which the negative impact on the quality of the simulations might be difficult to detect and quantify. Similarly, the transformation of unconventional observation types to more easily implementable observations could facilitate the use of unconventional observations in flow model calibration, but the transformation might rely on conceptual models or simplifications that are not appropriate and might result in model defects that are later

difficult to detect. Given the trade-offs of each implementation scheme for unconventional observations, the optimal choice on the implementation scheme should reflect the available information about the underlying conceptual system, the available observations, and the available time and resources (Anderson et al., 2015; Doherty, 2015; Hrachowitz & Clark, 2017; Jakeman & Hornberger, 1993).

2. **Challenges associated with the choice of the calibration strategy:** Which technique should be used to compare model outputs to unconventional observations, and how should the model parameters be calibrated? Should model parameters (i) be calibrated sequentially with multiple single-variable objective functions, calibrating one parameter (type) against one observation (type) at a time, or (ii) should all model parameters simultaneously be calibrated against all available observations using one single, weighted multivariate objective function? Sequential calibration of parameters is likely to result in a parametrization that favors an optimal solution related to the first observation (type) and first parameter (type) that are used, impairing an appropriate calibration of the subsequent parameters and observation types (see Townley, 2012). This problem can be overcome by simultaneously calibrating all model parameters against all available observations with one weighted multivariate objective function. However, finding an appropriate weighting scheme for a simultaneous calibration of all parameters against all available observations is often not straightforward and might result in a similar problem where the observation type with the highest associated weight is favored. If weighted appropriately, however, a simultaneous calibration of all model parameters against all available observation types has the potential to provide a better model parametrization (e.g., Anderson et al., 2015; Doherty & Hunt, 2010; Doherty & Welter, 2010; Hill & Tiedeman, 2007; Townley, 2012).
3. **Challenges associated with the ranking and weighting of observation types and individual observations**: Depending on the calibration order in case of a sequential calibration and on the weighting scheme in case of a simultaneous calibration, very different sets of calibrated parameters can result. Whether model parameters are calibrated sequentially or simultaneously, one needs to (i) rank observation types and parameters to decide which one to use first in a sequential calibration, and (ii) weight the different observation types for the use in a weighted single-variable or multivariate objective function. Ranking and weighting different observations and observation types, however, is not straightforward because observations can vary greatly in their number, in their measurement accuracy, and in their coverage of spatial and temporal scales. As a result of the imbalance of spatial and temporal scales of different observations, purely ranking/weighting observations according to their measurement accuracy (i.e., error-based weighting, which if the model were structurally perfect and the observations were perfectly balanced would be the mathematically optimal procedure; Hill & Tiedeman, 2007) does not always provide an adequate solution to the inverse problem (Doherty, 2015; Doherty & Welter, 2010). To account for the imbalances between different observations and for the fact that models are imperfect representations of reality, the initial error-based ranking or weighting are often adjusted to an interobservation-type-balanced weighting, guaranteeing that every observation type is similarly accounted for during calibration (see Anderson et al., 2015; Doherty & Hunt, 2010; Doherty & Welter, 2010; McCallum et al., 2012). As this procedure does not guarantee an appropriate model parametrization, and as no single perfect weighting scheme exists, the impact of multiple different weighting schemes should be assessed, and the worth of different observations in calibrating the model quantified (see Schilling et al., 2014). This allows assessing the robustness of conclusions under different weighting schemes. However, this procedure is laborious and therefore only rarely applied.

1.1 Aim, Strategy, and Structure of This Review

The specific objectives of this review are (1) to analyze the current use of unconventional observations in flow model calibration, (2) to identify the procedures that allow the most successful integration of unconventional observation types into flow model calibration and lead to better flow models, and (3) to investigate the information content or data worth of available unconventional observation types. The overall goal of this review is to identify optimal unconventional observation types and procedures for successful flow model calibration and to highlight current research needs.

The minimum requirement for studies to be considered in this review was for GW flow to be simulated in a physically-based way and on a regional scale (i.e., at least on the scale of a typical GW contaminant plume). Models ranging from physically-based, pure GW flow models to physically-based GW flow models including SW flow (i.e., an IFM), soil processes, or atmospheric processes were considered. Moreover, both pure flow models and flow models in combination with heat, energy, or mass transport were considered. One exception to these rules is a plot-scale SW-GW flow study, which was included due to the rigorous implementation of data worth quantification of different unconventional observation types.

In order to recapitulate the current understanding of how flow models should be calibrated, a brief tutorial on automated flow model calibration, uncertainty quantification, and analysis of the information content of different observation types is provided. The review itself is divided into subsections, each covering a different observation type. Each subsection begins with a brief review of relevant measurement techniques, measurement accuracy, and measurement errors; is followed by a detailed review of the studies that included the respective observation type in flow model calibration; and finishes with observation-type-specific conclusions. The review is then completed with an overall summary and discussion.

2 Tutorial: Automated Flow Model Calibration and Data Worth Analysis

To find an optimal parameter set in the context of flow model calibration, many different mathematical approaches, that is, many different automated inverse models, exist. Throughout the past decades, inverse methods have been the subject of multiple systematic comparisons (Carrera et al., 2005; de Marsily et al., 2000; Hendricks Franssen et al., 2009; McLaughlin & Townley, 1996; Rajabi et al., 2018; Townley & Wilson, 1985; W. W.-G. Yeh, 1986; Zhou et al., 2014). An excellent overview and classification over the many existing frequentist- and Bayesian-based inverse methods were recently provided by Rajabi et al. (2018). While the available inverse methods differ substantially in their underlying concepts and algorithms, it was found that differences among the resulting calibrated parameter sets are comparably small (Gallagher & Doherty, 2007; Hendricks Franssen et al., 2009), whether they are obtained with frequentist methods such as maximum likelihood estimation (MLE) based on the Gauss-Marquardt-Levenberg algorithm or with Bayesian methods such as ensemble Kalman filter (EnKF; Evensen, 1994; Hendricks Franssen & Kinzelbach, 2008; Kurtz et al., 2017) or Markov Chain Monte Carlo (Vrugt, 2016; Vrugt & Beven, 2018). While the mathematical approaches differ, the literature agrees that the calibration of flow models should be based on an automated inversion during which the mismatch between simulated and observed values is minimized according to a sound mathematical framework, rather than on manual trial-and-error calibration (e.g., Anderson et al., 2015; Doherty, 2015; Doherty & Hunt, 2010; Hill & Tiedeman, 2007; Poeter & Hill, 1997; Townley, 2012). Due to the strong complexity and dynamics of complex SW-GW flow systems,

complete protection against finding biased parameters even with mathematically sound and automated flow model calibration tools does not exist, be it frequentist or Bayesian based. However, mathematically sound and automated flow model calibration tools help in avoiding the local minima solution that are often found in sequential, manual trial-and-error calibration and allow the quantification of predictive uncertainty and of data worth in a mathematically-robust way even for complex IFMs. Automated flow model calibration that is based on a weighted multivariate objective function is, moreover, highly suited for simultaneous calibration against many different observation types and for the consideration all model features that can contribute to uncertainty.

As making predictions is most often a central purpose of flow models, systematically quantifying the uncertainty of predictions forms a key aspect of flow model evaluation (Anderson et al., 2015; Dausman et al., 2010; Delottier et al., 2016; Doherty, 2015; Hill & Tiedeman, 2007; Hunt et al., 2006; Jakeman et al., 2016; Schilling et al., 2014; Townley, 2012; White et al., 2016). The worth of different observations/observation types in reducing the uncertainty of flow model predictions provides an important piece of information for an optimal design of measurement campaigns and flow model calibration. The information content, or data worth, is of particular interest when using unconventional observations for flow model calibration. Automated inversion based on multivariate objective functions not only allows calibrating flow models against many different observation types simultaneously with a sound mathematical framework but also allows systematic quantification of the information content of the observations that are used during calibration—key information for optimal planning of measurement campaigns.

While the cost of complete post-palibration uncertainty analysis is high, several efficient albeit approximate uncertainty quantification techniques, which take advantage of the sensitivity matrices calculated during automated calibration, exist (e.g., Doherty, 2015; White et al., 2016). We here provide a brief tutorial on the design of automated flow model calibration and on the analysis of the information content of different observation types.

2.1 Weighted Multivariate Objective Function: Simultaneous Calibration of All Model Parameters Against All Available Observations

Two philosophies for automated flow model calibration exist: (i) sequential calibration of individual parameters or selected parameter groups against individual observation types (e.g., first hydraulic conductivity against hydraulic heads and then porosity against tracer concentrations) and (ii) simultaneous calibration of all parameters against all available observations. While it is indeed possible to calibrate the parameters of a flow model sequentially using a separate, weighted single-variable objective function per parameter-observation pairing, this procedure strongly relies on subjective judgment in the choice of the calibration sequence and is prone to result in calibrated parameters and predictions, which are dominated by the first parameter-observation pairing that was used during calibration (Doherty, 2015; Hill & Tiedeman, 2007; Townley, 2012). To avoid the problem of the calibration being dominated by one parameter group and one observation type, the calibration of flow models should be undertaken in a simultaneous calibration routine, in which all parameters are simultaneously calibrated against all available observation types using a weighted multivariate objective function (Doherty, 2015; Doherty & Hunt, 2010; Hill & Tiedeman, 2007; McLaughlin & Townley, 1996; Townley, 2012). Another positive aspect of simultaneous calibration is the fact that each parameter can be informed by multiple different observations and observation types at the same time, allowing information to flow more freely between observations and parameters. With the strong increase in affordable computational resources during the last two decades, especially owing to the

development of computational superclusters and cloud infrastructure, automated and simultaneous calibration of flow models, even with hundreds of parameters, has now made sequential flow model calibration obsolete (e.g., Delottier et al., 2016; Hunt et al., 2010; Kurtz et al., 2017). Unfortunately, sequential flow model calibration, and even manual trial-and-error calibration, is still applied often, despite the large potential for biased results, the widely available computational resources, and the many available automated flow model calibration algorithms.

Many different ways of formulating the GW inverse problem exist (Anderson et al., 2015; Moriasi et al., 2017). For simultaneous flow model calibration against two different observation types a and b (e.g., hydraulic heads and tracer concentrations), the GW inverse problem is commonly formulated as a weighted multivariate least squares objective function according to

$$\Phi = \sum_{i=1}^{n} \left[w_i (a_o - a_s)_i^2 \right] + \sum_{j=1}^{m} \left[w_j (b_o - b_s)_j^2 \right], \tag{1}$$

where

Φ is the objective function to be minimized,

$a_{o,i}$ and $b_{o,j}$ are ith observation of a and jth observation of b,

$a_{s,i}$ and $b_{s,j}$ are the simulated counterparts of $a_{o,i}$ and $b_{o,j}$,

w_i and w_j are the weights given to the ith and jth observation, and

n and m are the number of observations of observation type a_o and b_o, respectively.

A crucial aspect of using weighted multivariate objective functions is determining the weighting scheme to define w. In weighted single-variable objective functions used in sequential calibration, the weight w given to each observation should reflect the inverse of its measurement uncertainty in order to normalize observations of the same observation type by their accuracy (see Hill & Tiedeman, 2007). In weighted multivariate objective functions of simultaneous calibration, however, this weighting scheme often needs to be adjusted—for example, to avoid the calibration being dominated by the observation type with the largest number of observations (which in the GW context typically is hydraulic head). No single best weighting scheme exists, and the weighting scheme for calibration based on a weighted multivariate objective function needs to be chosen with the available observations and the purpose of the model in mind (Anderson et al., 2015; Doherty & Hunt, 2010; Doherty & Welter, 2010). If, for example, predictions of spring discharge are of interest, spring discharge observations should be weighted so that they make up an important fraction of the multivariate objective function. If spring discharge observations are much fewer in number than hydraulic head observations, the weight of spring discharge observations may need to be increased in order to guarantee their visibility in the multivariate objective function. In many cases, a suitable weighting scheme consists of balancing the different observation types so that they each have approximately equal impact on the multivariate objective function (Doherty & Hunt, 2010). However, many different opinions exist on the matter, and rather than strictly following one particular weighting scheme, it is more important that the applied weights and the reasoning behind them are well described in each study, so that readers can understand what influence each observation type has had on the outcome of flow model calibration.

Another important feature that should be included in automated flow model calibration is regularization, as it provides a mathematically sound way of dealing with the problem of parameter nonuniqueness, that is, the problem that many different equally likely solutions exist to the flow model inverse problem (Doherty, 2015; Doherty & Hunt, 2010; Moore et al., 2010; Moore & Doherty, 2006). Regularization creates uniqueness by specifying extra criteria for the parameters that are subject to calibration. However, it does not eliminate predictive uncertainty, as uniqueness is only artificially created, be it through some form of averaging to homogeneous values of otherwise heterogeneous properties or through other forms of regularization (e.g., by tying multiple parameters together by a fixed relationship). Nevertheless, there are many benefits of regularization. A robust regularization method is Tikhonov regularization, which provides a receptacle for direct information (or prior information or expert knowledge) on the parameters to be calibrated by introducing a penalty to the objective function if calibrated parameter values stray too far from the preferred values or preferred parameter relationships (such as differences or ratios; Delottier et al., 2016; Doherty, 2015; Doherty & Hunt, 2010; Moore & Doherty, 2006; Tikhonov & Arsenin, 1977). Tikhonov regularization provides a fallback value for parameters on which there is not enough information in the observation data set. This creates uniqueness where one would have nonuniqueness, a problem often encountered when only limited observations are available, or in models with large numbers of parameters that need to be calibrated. The direct integration of direct or prior information on parameters into the objective function also enables a mathematically sound quantification of the impact of this kind of information on the calibrated parameters. A simple Tikhonov regularization penalty function can be added to a weighted multivariate objective function with two observation types according to

$$\Phi = \sum_{i=1}^{n}\left[w_i(a_o - a_s)_i^2\right] + \sum_{j=1}^{m}\left[w_j(b_o - b_s)_j^2\right] + \sum_{k=1}^{q}\left[w_k(p_o - p_c)_k^2\right], \quad (2)$$

where

$p_{o,k}$ is kth preferred parameter value,

$p_{c,k}$ is kth calibrated parameter value,

w_k is the weight given to kth preferred parameter value, and

q is the number of parameters.

Assistance and more detailed information on the setup of automated flow model calibration and the implementation of Tikhonov regularization can, for example, be found in the documentation of the calibration software PEST (Doherty, 2015, 2016; Doherty et al., 2010).

2.2 Quantifying the Reduction of Predictive Uncertainty and the Worth of Different Observation Types

Due to the strong spatial and temporal complexity of environmental systems and our inability to measure this complexity, calibrated flow models remain only approximate representations of reality (Anderson et al., 2015; Haitjema, 2015). Consequently, it is impossible to quantify the true uncertainty of model predictions.

While it is impossible to quantify the true uncertainty of flow model predictions, the reduction in the predictive uncertainty that is achieved through calibration against observations can nevertheless be estimated with Bayesian statistics: Parameters and predictions of the uncalibrated model represent the prior probability distribution, and the calibrated model parameters and predictions represent the posterior probability distribution. The observations used for calibration are the events that help to reduce the predictive uncertainty of the flow model by constraining model parameters. Unfortunately, flow models can be computationally very demanding, and, consequently, quantifying the uncertainty of a flow model prediction can rarely be based on pure Bayesian statistics: A true Bayesian sampling of the prior probability distribution of model parameters and predictions, for example, through Monte Carlo or Latin Hypercube sampling, in most situations requires too many simulations to be carried out (see Doherty, 2015, or Hill and Tiedeman, 2007). The reduction of uncertainty of a model prediction can nevertheless be estimated with computationally efficient yet approximate approaches, for example, through predictive uncertainty analysis (Christensen & Doherty, 2008; Doherty, 2015; Doherty & Hunt, 2009; White et al., 2016) or polynomial chaos expansion (Miller et al., 2017). Owing to improved computational resources, also stochastic, Bayesian-based methods may be applied for predictive uncertainty quantification with GW flow models, for example, using the EnKF (Evensen, 1994; Hendricks Franssen & Kinzelbach, 2008; Kurtz et al., 2017), the iterative ensemble smoother (Chen & Oliver, 2012; White, 2018), or the Markov chain Monte Carlo method (Vrugt, 2016; Vrugt & Beven, 2018). The improvement of computational technology in recent years has enabled Bayesian methods to be applied even to relatively complex flow models, but conducting data worth analyses of different observation types with Bayesian methods for complex IFMs is computationally so demanding that it is unfeasible in most situations (Dai et al., 2016; Lu et al., 2018). A beneficial by-product of MLE based on the Gauss-Marquardt-Levenberg algorithm is that the data worth of different observation types can be efficiently assessed based on the calculated Jacobi matrix of sensitivities between model outputs and model parameters, without the need to rerun the calibration routine for different combinations of observation types over and over, as would be the case for Bayesian methods (Christensen & Doherty, 2008; Doherty, 2015; Doherty & Hunt, 2009; White et al., 2016). While Bayesian data worth quantification methods are constantly being improved (Dai et al., 2016; Lu et al., 2018) and computational resources are ever-increasing, an existing and computationally efficient way of data worth analysis based on the Jacobi matrix of sensitivities is presented here:

Data worth analysis based on the Jacobi matrix can be undertaken without much computational effort using Bayesian statistics but with the simplifying assumptions of linear model behavior, Gaussian distributions for prior parameter uncertainty, and Gaussian distributions of measurement errors (see Brunner et al., 2012; Doherty, 2015; Schilling et al., 2014). The benefits and the problems that are associated with these assumptions when evaluating model predictive uncertainty and data worth were discussed and compared to the nonlinear and more robust but computationally much more demanding Markov chain Monte Carlo method by Gallagher and Doherty (2007). Sensitivity analyses (see Hill & Tiedeman, 2007; Gannett et al., 2012; Masbruch et al., 2014; Rasa et al., 2013), influence statistics (see Hill & Tiedeman, 2007; La Vigna et al., 2016), and singular-value-decomposition-based principal component analysis (see Doherty, 2015; Doherty & Hunt, 2009; Schilling et al., 2014) provide additional computationally efficient means to estimate the information content of specific observations or observation types. All these methods are based on the sensitivity matrices calculated with automated flow model calibration routines that are based on weighted multivariate objective functions. Describing the details of all these methods is beyond the scope of this tutorial, but details of their implementation can, for example, be found in Doherty (2015), White et al. (2016), Hill and Tiedeman (2007), and the manuals to the calibration tools PEST and UCODE. Here details

on a linear predictive uncertainty reduction and data worth analysis that can be carried out with the PEST software suite GENLINPRED (Christensen & Doherty, 2008; Doherty, 2015, 2016) or with pyEMU (White et al., 2016)) and which was employed by Brunner et al. (2012), Dausman et al. (2010), Engelhardt et al. (2013), Fienen et al. (2010), and Schilling et al. (2014), are provided.

With PEST and GENLINPRED, the uncertainty σ_s^2 of a prediction s is calculated using the following equation presented by Christensen and Doherty (2008):

$$\sigma_s^2 = \mathbf{y}^t C(\mathbf{p})\mathbf{y} - \mathbf{y}^t C(\mathbf{p})\mathbf{X}^t \left[\mathbf{X}C(\mathbf{p})\mathbf{X}^t + C(\boldsymbol{\varepsilon})\right]^{-1}\mathbf{X}C(\mathbf{p})\mathbf{y}, \quad (3)$$

where
$\mathbf{p}$ is a vector of model parameters,

$C(\mathbf{p})$ is the prior covariance matrix expressing parameter uncertainty,

$\boldsymbol{\varepsilon}$ is a vector whose elements represent noise associated with elements of the calibration data set,

$C(\boldsymbol{\varepsilon})$ is a covariance matrix of measurement noise,

$\mathbf{X}$ is the sensitivity matrix under calibration conditions (this denotes the sensitivity of each model output used in the calibration process to each adjustable parameter), and

$\mathbf{y}$ is a vector comprised of the sensitivities of the prediction s to all model parameters.

The first term on the right-hand side of equation 3 provides the prior (i.e., pre-calibration) uncertainty of the prediction. The second term provides the reduction in this uncertainty gained through model calibration and serves thus as an approximate measure of the data worth of the calibration data set. The data worth of individual observations, or of an observation group, can be evaluated by calculating the uncertainty of a prediction s of interest with and without that observation or observation type included in equation 3. The reduction in uncertainty of the prediction gained through inclusion of that data in the calibration data set is a measure of its worth. By comparing σ_s^2 for different observation combinations, the method also allows identification of observation types that do not improve predictive uncertainty or may even deteriorate it.

3 Review of the Use of Unconventional Observation Types

An overview of the reviewed studies is provided in Table 2: Information about the study type, the employed observation types, the simulated processes, the modeling codes, the horizontal scale, and the calibration and uncertainty estimation methods is summarized. As outlined in section 1.1, this review is grouped by the way unconventional observations were used in the calibration procedure of flow models. As an example, despite the fact that TT and RT observations are commonly obtained by transforming tracer concentration observations, TT and RTs are discussed in their own dedicated section (section 3.5) rather than in the section on the use of tracer concentrations (section 3.4).

Table 2. Overview of the Reviewed Studies; the Models Used; the Calibration Techniques Applied; and of Which, as Well as How, Unconventional Observation Types Were Included

Study	Model type	Scale (km^2)	Additional processes	Process integration	Numerical model	Inverse model	Calib tech
Ala-Aho et al. (2015)	3-D real world	200			HydroGeoSphere	manual	sequ
Alaghmand et al. (2014)	3-D real world	1	mass	explicit	HydroGeoSphere	manual	sequ
Alaghmand et al. (2016)	3-D real world	0.5	mass	explicit	HydroGeoSphere	manual	sequ
Bartsch et al. (2014)	2-D real world	<0.001	heat	explicit	HydroGeoSphere	MLE (PEST)	simu
Bauer et al. (2001)	2-D real world	10	mass	explicit	MODFLOW + MT3DMS	manual	sequ
Bonton et al. (2012)	3-D real world	15	mass	explicit	HydroGeoSphere + AgriFlux	manual	sequ
Boronina, Balderer, et al.	2-D real world	300	mass	simplified	MODFLOW + PMPATH	manual	sequ

Note. Observation types: H = hydraulic heads, K = hydraulic conductivity/transmissivity, Q_{SW} = SW discharge, Ex = exchange fluxes, Q_{GW} = GW discharge, Q_{base} = baseflow, R = recharge, T = temperature, C = tracer concentration, RT/TT = residence/travel times, ET = evapotranspiration, S = soil moisture. Inverse models: manual = manual trial-and-error, MLE = maximum likelihood estimation (least squares regression), DA = Data assimilation, SSC = sequential self-calibration, EnKF = ensemble Kalman filter. Flow models: CATHY (Camporese et al., 2010), COMSOL (Q. Li et al., 2009), FEFLOW (Dirsch, 2014), FEMWATER (Yeh, 1987), HST3D (Kipp, 1997), HydroGeoSphere (Therrien et al., 2018; Kurtz et al., 2017), METIS (Goblet, 1999), MIKE-SHE/MIKE11 (Abbott et al., 1986; Graham & Butts, 2005; DHI, 2007), MOC (Konikow & Bredehoeft, 1978), MODFLOW (Harbaugh, 2005), SEAWAT (Langevin, 2009), SPRING (Delta h Ingenieurgesellschaft mbH, 2006), VS2DHI (Hsieh et al., 1999). Packages: AgriFlux (Banton & Larocque, 1997), HMC (Partington et al., 2011), MODFLOWT (Duffield et al., 1996), MODPATH (Pollock, 2016), MT3DMS (Zheng, 2010), PMPATH (Chiang & Kinzelbach, 2001), RT3D (Clement, 1997), SWAT (Bailey et al., 2016).

3.1 Temperature

3.1.1 Observing Temperature

Measurements of water temperature (T) are easy to obtain, can be made in almost every environment, and are typically available at a high precision (Anderson, 2005). Recent advances in airborne and fiber optic (FO) techniques have enabled distributed temperature sensing (DTS). In the case of FO-based DTS, for example,

temperature changes can be recorded continuously in space and in time (Shanafield et al., 2018; Vogt et al., 2010). Temperature measured with airborne hyperspectral and near-infrared cameras, on the other hand, allows for the generation of detailed maps of surface temperature, albeit at lower temporal resolution than with FO-DTS (Cardenas et al., 2011). Temperature contains information on both flow and conduction of heat, as well as on temperature sources and sinks. For these reasons, temperature measurements have received a lot of attention as a SW-GW tracer and as an indicator for exchange fluxes and mixing processes (e.g., Anderson, 2005; Bravo et al., 2002; Harvey & Gooseff, 2015; Hunt et al., 1996; Johnson et al., 2005; Shanafield & Cook, 2014).

However, due to the many different processes that influence temperature, the explicit simulation of heat transport requires a relatively large number of additional parameters and boundary conditions to be defined (see Table 1). Moreover, the information contained in observations of temperature is often confounded with regard to flow due to the multitude of different processes that influence the spatial and temporal dynamics of T. This makes the information content of observations of T limited to the time and location of the measurement, as demonstrated, for example, by Bravo et al. (2002), Shanafield and Cook (2014), Irvine et al. (2015), and Xie et al. (2015). Whether a flow model that is calibrated against a combination of classical observations and observations of T alone is able to simulate water fluxes and heat and mass transport is unclear.

3.1.2 Applications of Temperature Observations

In one of the earliest and most thorough attempts to reduce the nonuniqueness of flow models through automated calibration with a weighted multivariate objective function, Bravo et al. (2002) calibrated multiple synthetic models and one real-world flow and heat transport model of wetland-aquifer systems against observations of H and T. Through a first synthetic example, Bravo et al. (2002) demonstrated that if the thermal properties and the effective porosity f of a simple 1-D vertical and homogeneous porous medium model are perfectly known, calibration of K against observations of H and T allows accurate determination of K. Through a second synthetic example of a 2-D cross-sectional and homogeneous wetland-aquifer model, they showed that K could be inferred correctly from a calibration against H if thermal parameters f and the exchange flux between the wetland and the aquifer are perfectly known. When both K and the exchange flux were unknown and needed calibration, however, they could only be estimated through a combination of both H and T. As in both of these two initial synthetic examples, the synthetic observations of H and T used for calibration did not contain any measurement noise and the porous medium was homogeneous, Bravo et al. (2002) carried out a third synthetic example with a 2-D cross-sectional, two-layer wetland-aquifer model, in which they also added measurement noise to the synthetic observations. In this case, a simultaneous estimation of K and exchange fluxes between the wetland and the aquifer was not possible unless the initial guesses were already sufficiently close to the correct values, clearly showing that the inverse problem is ill-posed. The authors subsequently conducted a calibration study of a 3-D real-world wetland-aquifer flow and heat transport model. They compared calibrated K to measured point values obtained through pumping tests, as well as simulated exchange fluxes to field-based estimates of exchange fluxes. They were able to show that the inclusion of observations of T allowed the observed K values to be more closely matched than by calibration against observations of H alone. They also demonstrated that simultaneously calibrating the model against observations of H and T resulted in exchange fluxes similar in order of magnitude to the field-based exchange flux estimates. However, due to the synthetic findings and the confounded nature of T, Bravo et al. (2002) pointed out that including observations of T to calibrate hydraulic parameters is highly likely to result in biased results, specifically highlighting the fact that the geological stratigraphy of thermal

properties does not necessarily coincide with the geological stratigraphy of hydraulic properties. Unfortunately, many of the later studies that incorporated observations of T alongside classical observations did not conduct such rigorous model calibration and model validation routines.

Manning and Solomon (2005), Mutiti and Levy (2010), and Heilweil et al. (2012) only calibrated their IFMs by manual trial-and-error procedures. They nevertheless concluded that they were able to improve their model parametrizations by considering observations of T alongside classical ones: Manning and Solomon (2005) considered direct observations of T alongside observations of noble gas recharge temperatures (NGRTs), GW discharge (Q_{GW}) and RTs. Based on the comparison between simulations with an IFM of coupled flow, heat and mass transport, and the available observations, they distinguished adequate flow regimes from inadequate ones. The authors stated that observations of T were a key observation type for the identification of inadequate flow regimes. Similarly, Heilweil et al. (2012) manually calibrated a cross-sectional flow model of a volcanic island against observations of H, T, and of NGRTs, and showed that only the comparison of simulated to observed T allowed recharge locations to be constrained and aquifer K (K_{aq}) to match all observation types simultaneously. Mutiti and Levy (2010) used observations of T alongside observations of H to reproduce the transient changes in K of the riverbed (K_{rb}) resulting from large storm runoff events. Manually calibrating K_{rb} against observations of H alone did not allow the transient temperature changes observed during and after a storm runoff event to be reproduced. According to Mutiti and Levy (2010), however, the addition of observations of T to the manual calibration allowed adapting K_{rb}, which resulted in a better reproduction of observed temperature patterns. The authors did not compare these results to observations of exchange fluxes and could therefore not determine if the improved representation of temperature patterns coincided with an improved representation of water fluxes.

Moving a step forward from manual to automated calibration, Naranjo et al. (2012) investigated the identifiability of hydraulic and thermal parameters on the basis of a real-world cross-sectional river-aquifer model that they calibrated against observations of H and T with a uniform random sampling approach. The central aim of their model was to understand hyporheic flow paths. Simulations based on the parameters that were identified as the most appropriate through the stochastic sampling approach allowed Naranjo et al. (2012) to reproduce the observed H and T with their IFM. The comparison between the measured shallow K_{aq} and the calibrated K_{aq}, however, revealed a significant mismatch, and hyporheic exchange fluxes were not compared to direct flux observations. Also, Bartsch et al. (2014) calibrated a real-world cross-sectional river-aquifer flow model against observations of H and T. Calibration was based on a weighted multivariate objective function as implemented in PEST, and the aim of the calibrated flow model was to characterize the influence of monsoonal-climate-controlled river-aquifer exchange fluxes on the transport of nutrients and contaminants. Like the calibrated models of Mutiti and Levy (2010) and Naranjo et al. (2012), the calibrated model of Bartsch et al. (2014) also was able to reproduce the observed H and T well. However, neither Bartsch et al. (2014) directly compared flux simulations to flux observations, and the conclusion on whether the model was suited for water flux and mass transport simulations was only based on comparison of simulated to observed H and T, and on comparison of textural-soil-data-derived K_{aq} to the calibrated K_{aq}.

Munz et al. (2017) conducted a sophisticated water flow and heat transport study of a real-world river-aquifer system using a 3-D IFM model, calibrated against a large number of observations of H and T. The authors calibrated multiple thermal and hydraulic parameters, including separate heterogeneous K_{rb}- and K_{aq}-fields, using the automated, weighted multivariate objective function calibration routine implemented in PEST. An extensive global sensitivity analysis based on 780 model runs was carried out prior to model calibration using Morris' elementary effects screening method (Morris, 1991). A subsequent calibration was focusing on the

parameters most sensitive to observations of H and T. The sensitivity analysis revealed that (1) anisotropy of the hydraulic conductivity and density of the bulk material were sensitive to observations of H; (2) the hydraulic conductivity mean, variance, and anisotropy, as well as the density of the bulk material and the thermal conductivity of the sediment, were sensitive to observations of GW temperatures; and that (3) the hydraulic conductivity range and anisotropy, as well as the density of the bulk material, the thermal conductivity of the sediment, and the heat capacity of the sediment were sensitive to observations of riverbed temperatures. The calibrated model was able to reproduce the observed H and T well. Munz et al. (2017) used the model to investigate the temperature distributions within the hyporheic zone and the aquifer and found that the water temperature is strongly dependent on the RT of water in the hyporheic zone. The purpose of the model was not to analyze the magnitude of water fluxes. Consequently, the model was not compared to flux observations, and the data worth of observations of T for the reproduction of water fluxes was not assessed. Munz et al. (2017) noted that the calibrated mean horizontal hydraulic conductivity was 1.6 times higher than measured values. Like Munz et al. (2017), Karan et al. (2014) also calibrated K_{rb} of a 3-D river-aquifer model against observations of H and T based on automated calibration with a multivariate objective function using PEST. Trying to limit the confounding nature of T observations for the reproduction of fluxes, Karan et al. (2014) employed an exceptionally high-resolution measurement network of 400 temperature sensors within a 10 × 10 × 0.5 m riverbed section and calibrated three structurally different models. Owing to the very high resolution of T observations, they could show that through the inclusion of T alongside H the structurally inappropriate models (i.e., a homogeneous and a two-zone model) could be distinguished from the appropriate one (i.e., heterogeneous K_{rb} and K_{aq}). Unfortunately, the simulated exchange fluxes were not validated with measured exchange fluxes, making it difficult to assess the data worth of observations of T beyond that for the reproduction of hydraulic heads and temperature. Shope et al. (2012) also calibrated a 3-D, steady state flow and heat transport model of a highly instrumented gravel bar section within a larger river-aquifer system against observations of H and T, aiming to understand the influence of gravel bars on SW-GW exchange fluxes and water table dynamics. The model was calibrated using an automated calibration routine that was based on a multivariate objective function. Owing to high-resolution GW temperature measurements down to 2-m depth within the gravel bar, Shope et al. (2012) were able to simultaneously match H and T well. The match between simulated and observed temperatures in the river banks, which were measured at depths ranging from 4 to 11 m, was significantly worse. Unfortunately, no information on the weighting scheme was provided, and no uncertainty or data worth analyses were carried out.

After Bravo et al. (2002), it took many years until new studies were published, which continued to systematically investigate the confounding nature of T observations in the context of reproducing fluxes: In a rigorous IFM calibration study, Kurtz et al. (2014) simultaneously assimilated observations of H and T in both a synthetic and a real-world IFM of an alluvial drinking water station using the EnKF approach (Evensen, 1994; Hendricks Franssen & Kinzelbach, 2008; Kurtz et al., 2017; Tang et al., 2017). They found that the coupled state and parameter updating through assimilation of H and T improved the quality of the simulations compared to an assimilation of H alone. However, they also found that assimilation of H and T only resulted in an improved characterization of the hydraulic heads and the temperature, but not of the flow field. Adding additional unconventional observation types to the calibration data set, Xu and Gómez-Hernández (2016) applied an EnKF approach and simultaneously assimilated observations of H and T combined with observations of tracer concentrations (C). They calibrated K_{aq} and f in a synthetic modeling study and could show that the best result was obtained by using all three different observation types in combination. However, in their synthetic study the thermal properties of the subsurface were already perfectly

known prior to calibration, and only K_{aq} and f were uncertain; major confounding effects on temperature were therefore not present. Ma et al. (2012) compared the benefit of using observations of H, T, and of C in a calibration of a heat and mass transport model using a weighted multivariate objective function and showed, like Bravo et al. (2002), that confounding effects of T are a problem for the reproduction of flow in a real-world scenario. It was found that observations of T are inferior compared to observations of C for inferring parameters of flow. Ma et al. (2012), however, showed that if observations of T can be obtained under favorable conditions, that is, under mainly advection-controlled heat transport with well-known sources and sinks, observations of T can be beneficial at locations where observations of C are too difficult or costly to obtain. Unfortunately, no information on the weights applied to the different observations in the calibration objective function were provided.

Engelhardt et al. (2013) came to a different conclusion in terms of the information content of observations of T versus observations of C for the reproduction of fluxes: They investigated the information content of observations of H, T, and C (concentrations of acesulfame) for the calibration of a real-world cross-sectional river-aquifer model. They used PEST to simultaneously calibrate a coupled flow, heat, and mass transport model against observations of H, T, and C with a weighted multivariate objective function and subsequently quantified data worth through a linear predictive uncertainty analysis. The calibrated model was able to reproduce both the observed H and T well, and to a lesser degree also the observed C. Through the data worth analysis, Engelhardt et al. (2013) could demonstrate that hydraulic and thermal parameters were both sensitive to observations of T and that, compared to observations of C, observations of H and T were of higher data worth for the reduction of the uncertainty of mass transport and exchange flux predictions. Similar to the finding of Munz et al. (2017), water temperatures at shallow depths directly underneath the river were identified as the most crucial information for exchange flux and transport simulations, and Engelhardt et al. (2013) concluded that this subset of observations of T is more important for mass transport simulations compared to direct observations of C. However, Engelhardt et al. (2013) noted that the observed C of acesulfame was almost constant throughout the entire system, whereas T was strongly variable, and that this could be the reason for the higher data worth of observations of T compared to C for the reproduction of fluxes. It is also worth noting that observations of H and T were available in much larger numbers compared to observations of C. From the reported details on the calibration and data worth analyses, it is not entirely clear whether the three different observation types were adjusted to feature equally in the multivariate objective function, or whether the total weights of H, T, and C were all different and potentially contributing to the differences in data worth.

Adding yet more observation types to model calibration and information content analysis, Masbruch et al. (2014) calibrated a 3-D coupled flow, heat, and transport model of a real-world system against observations of H, Q_{SW}, T, and exchange fluxes (baseflow, spring discharge, and ET). Using the procedures implemented in UCODE, they employed a rigorous, multivariate-objective-function-based calibration and systematically assessed the sensitivities between, and the information content of, the different unconventional observation types toward the calibrated parameters (quantifying composite scaled sensitivities, parameter correlations, influence statistics, and posterior confidence intervals). The thorough analysis revealed that observations of T can help inform hydraulic parameters in the absence of spring discharge observations. However, as soon as observations of Q_{GW} in the form of spring discharge were included alongside observations of H, Q_{SW}, and T, the information from T for hydraulic parameters became negligible—observations of T then only helped informing thermal parameters. In another equally rigorous study that systematically analyzed the data worth of observations of T alongside other observation types, Delsmann et al. (2016) quantified the value of

observations of H, T, Q_{GW}, and C for the calibration of an IFM. The worth of different observation types in calibrating an IFM was estimated using a uniform random sampling approach (generalized likelihood uncertainty estimation; Beven & Binley, 1992; Stedinger et al., 2008). Like Bravo et al. (2002) and Kurtz et al. (2014), Delsmann et al. (2016) carried out both synthetic and real-world modeling experiments to investigate the data worth of different observation types for flow model calibration. Delsmann et al. (2016) showed that, overall, the inclusion of unconventional observations improved model parameters and predictions if the observations were of a sufficient quality, that is, if the observations were associated with a sufficiently small measurement error. They found, however, that with observations of T only two parameters could be calibrated well: f and the thermal conductivity of the porous medium (κ_{bulk}). Delsmann et al. (2016) also found that (1) all other observation types employed (i.e., H, Q_{GW}, and C) constrained the hydraulic gradients and flow field better than observations of T and that (2) by including observations of T, all model outcomes other than the simulation of heat transport worsened. Only the inclusion of observations of C (in the form of total salinity load) allowed simultaneously improving all model outcomes.

3.1.3 Conclusions on Using Temperature Observations

The inclusion of observations of T alongside classical observations is clearly beneficial for the simulation of heat transport—this has been demonstrated in all studies reviewed above. The studies showed that when only classical observations of H were used for the calibration of flow and heat transport in a flow model, heat transport could not be adequately simulated. However, the application of observations of T for the calibration of a model aimed at simulating exchange fluxes and mass transport is associated with many pitfalls and potentially negative impacts on all model outcomes other than those of heat transport. The review showed that more often than not, the validity of flux or mass transport simulations calibrated against a combination of classical observations and T was not assessed by directly comparing the simulations to flux or mass transport observations. This is not good practice, especially in the light of the findings from the thorough analyses carried out by Bravo et al. (2002), Kurtz et al. (2014), Masbruch et al. (2014), and Delsmann et al. (2016), who demonstrated that as soon as there are multiple unknowns among the hydraulic and thermal properties and relevant exchange fluxes, and as soon as measurement errors are taken into account, calibration of an IFM against H and T is likely to result in correct temperature reproduction but incorrect representation of fluxes. If observations of T were complemented by a combination of classical and additional unconventional observations (e.g., by C or Q_{GW}), an acceptable calibration of flow, heat, and mass transport was much more likely. Following from the findings of Shope et al. (2012), Engelhardt et al. (2013), Karan et al. (2014), and Munz et al. (2017), the highest information content of temperature observations for the reproduction of fluxes can be attributed to temperature observations at shallow depths up to 2 m directly underneath a river, where they help inform exchange fluxes between the river and the underlying aquifer.

The findings of the studies reviewed above match well the findings of Irvine et al. (2015), who compared the suitability of observations of T to the suitability of observations of C as environmental tracers in heterogeneous media: Irvine et al. (2015) concluded that observations of T are more influenced by mixing than C, which makes T a better tracer for the average flow velocity but less suited for an analysis of the spatial distribution of flow. Thus, if measurements of temperature are taken only 1–2 m below the river, they may still carry some information about the spatial distribution of exchange fluxes, but this information is lost in measurements of T taken at larger depths. Similarly, Xie et al. (2015) and Shanafield and Cook (2014) found that temperature observations directly taken in the SW do not contain sufficient information for the quantification of SW-GW exchange flux patterns.

The practice of using temperature observations for the calibration of deeper GW flow systems is surprising in the light of the many findings reviewed above, even more so as the confounding nature of observations of T has already been discussed in-depth by Bravo et al. (2002) early on in the history of flow model calibration against unconventional observations: Bravo et al. (2002) highlighted almost two decades ago that observations of T are confounded, as the hydraulically relevant subsurface structure might differ substantially from the thermally relevant subsurface structure, and observations of T should therefore only be used to calibrate hydraulic parameters where direct observations of water fluxes are unavailable. The reviewed studies on the use of observations of T also revealed another crucial aspect of flow model calibration studies: Many different calibration procedures and weighting/ranking schemes were applied, and generally, only limited information on which procedure was used and how observations were weighted was provided (see Table 2).

3.2 Exchange Fluxes

3.2.1 Observing Exchange Fluxes

Exchange fluxes between GW, SW, and the atmosphere include

- diffuse recharge (R),
- infiltration from a SW body into the subsurface (I),
- discharge of GW into a SW body,
  - baseflow to rivers (Q_{base}),
  - diffuse and spring discharge (Q_{GW}),
- ET.

Observations of exchange fluxes are more difficult to obtain than classical hydrological observations. The easiest exchange flux observation to obtain is spring discharge (i.e., Q_{GW}), which can be measured directly. Local observations of diffuse Q_{GW} and I can also be obtained relatively easily with seepage meters. However, these observations typically provide only very local information. Owing to the rapid advances in remote sensing technologies, it is now possible to estimate R (or I) and ET on a much larger spatial and temporal scale. While remotely sensed R and ET are uncertain due to the fact that models are required to estimate R and ET from the remotely sensed data (e.g., Abtew & Melesse, 2013), the fact that remotely sensed R and ET cover such large spatial and temporal scales partially makes up for their limited accuracy. The value of remotely sensed observations of R and ET is particularly high for remote regions where the classical hydrological observations of H and Q_{SW} and local point measurements of exchange fluxes are scarce or completely absent. Observations of Q_{base} can be obtained by measuring the discharge of a river or stream during multiple extended low-flow conditions in the absence of recharge. However, as such periods are rare and as the influence of natural and anthropogenic mechanisms on Q_{SW} are complex, representative observations of Q_{base} are difficult to obtain and therefore often uncertain (Brunner et al., 2017; Partington et al., 2017; Smakhtin, 2001; Tallaksen, 1995).

Exchange flux measurements are often obtained indirectly: Diffuse discharge of GW into a SW body (Q_{GW}), infiltration of SW into GW (I), and recharge (R) are often quantified based on measurements of tracer concentrations (Cook, 2013; Harvey & Gooseff, 2015; Healy & Scanlon, 2010; McCallum et al., 2012; Shanafield & Cook, 2014), temperatures (Anderson, 2005; Healy & Scanlon, 2010), or soil moisture dynamics (Brunner et al., ; Healy & Scanlon,). However, as with remotely sensed exchange

2004 2010

fluxes, exchange fluxes based on such measurements are uncertain due to the transformations that are needed (Cook, 2013; Delottier et al., 2018; Healy & Scanlon, 2010; Shanafield & Cook, 2014; Xie et al., 2015). In this section, all studies that applied observations of exchange fluxes are discussed, no matter the measurement method employed. The studies in which observations of temperatures, soil moisture, and tracer concentrations were directly used for model calibration are discussed in the respective section on temperatures (section 3.1), on soil moisture (section 3.3), or on tracer concentrations (section 3.4).

The integration of exchange fluxes into IFMs is straightforward, as IFMs simulate exchange fluxes by default and therefore do not require exchange fluxes to be imposed as boundary conditions. Simulating exchange fluxes in pure GW models that do not simulate SW flow in a coupled and physically-based way is, on the other hand, associated with uncertainties, as the exchange fluxes need to be imposed on the model. Nevertheless, as the reviewed studies below have demonstrated, the consideration of exchange fluxes can even improve the calibration of pure GW models beyond that which can be achieved by considering only classical observations.

3.2.2 Applications of Exchange Flux Observations

Hunt et al. (2006) systematically assessed the value of diverse observation types in the calibration of a regional IFM in a rigorous study. Besides classical observations of H and lake stage, they also included unconventional observations of Q_{base}, Q_{GW} into a lake, I out of a lake, depth of a subsurface plume with elevated electrical conductivity (EC), and TTs in the calibration data set. The model was calibrated using a weighted multivariate objective function that simultaneously considered all observation types. Through a systematic data worth analysis based on influence statistics (Cook's D; Hill & Tiedeman, 2007), Hunt et al. (2006) found that spatially distributed observations of Q_{base} were the most informative observation type overall, allowing exchange fluxes to be allocated to different subbasins. I and Q_{GW} were the most important observation types to quantify lakebed leakage. In a similarly rigorous study, La Vigna et al. (2016) systematically assessed the information content of observations of H and Q_{GW} (spring discharge) in the calibration a regional SW-GW model with a weighted multivariate objective function. An analysis of composite scaled sensitivities, parameter correlation coefficients, and leverage statistics (see Hill & Tiedeman, 2007) revealed that the observations of Q_{GW} allowed not only parameters that are sensitive to observations of H to be informed but also parameters that are not sensitive to H; the information content of Q_{GW} was therefore essential in calibrating parameters that would otherwise have remained uninformed by classical observations. This was found even though only 3 data points of Q_{GW} were used alongside 13 data points of H, and despite the fact that the weighting scheme purely reflected the measurement uncertainty and was not scaled to guarantee equal importance to both observation groups. Also Masbruch et al. (2014), in a study already discussed in section 3.1 in relation to observations of T, showed with their data worth analysis that among all their observation types, the most informative for most of their model parameters was Q_{GW} from spring discharge. Observations of ET and Q_{base}, however, had negligible information content for their flow parameters if used alongside Q_{GW}.

That unconventional observations can be useful in regions where observations of H are scarce has also been demonstrated by Hendricks Franssen et al. (2008), who used spatially distributed R that was obtained through remotely sensed soil moisture measurements (see Brunner et al., 2004). The absolute magnitude and the spatial pattern of R were used as unconventional observation types alongside observations of H to stochastically generate hundreds of equally likely solutions to the inverse problem of a regional flow model.

The large uncertainty in the spatial distribution of R obtained through remote sensing and chloride methods was also considered. Hendricks Franssen et al. (2008) concluded that the inclusion of remotely sensed observations of absolute magnitudes and of the spatial pattern of R substantially improved the model outcome compared to a model that would have only been constrained by H. The highest data worth was related to spatial pattern information of R in areas where observations of H were scarce. Knowling and Werner (2016) were among the first to calibrate a regional IFM in order to investigate the identifiability of spatially distributed R and spatially distributed K_{aq} by inclusion of observations of Q_{GW} (in the form of subsurface GW discharge to the ocean) alongside observations of H. They showed that the inclusion of Q_{GW} did improve the overall estimation of R in the catchment; however, the estimation of R without direct observations of R proved to be difficult once they tried to simultaneously identify R and the heterogeneous K_{aq}-field. The combination of observations of Q_{GW} and of H thus did not contain sufficient information to calibrate the spatial distribution of R and K_{aq} simultaneously. Delsmann et al. (2016), whose study was already discussed in the context of observations of T (see section 3.1), also applied observations of Q_{GW} into tiles and a ditch alongside observations of H, T, and C. The inclusion of Q_{GW} substantially improved the overall reliability of the model, particularly when the calibrated model was compared to one calibrated only against classical observations of H. Ala-Aho et al. (2015) included observations of Q_{base} and Q_{GW} to rivers and a lake alongside classical observations of H and lake stage to calibrate a 3-D flow model of a regional SW-GW system. While their regional IFM was not calibrated using an automated procedure, choosing parameter values that matched all the different observations types simultaneously in a manual trial-and-error procedure allowed solutions that would have correctly reproduced H and lake stages but failed in reproducing Q_{GW} to lakes to be discarded. Manning and Solomon (2005) also applied multiple different observation types in a manual trial-and-error calibration and could distinguish probable from impossible flow regimes of a large mountainous catchment mainly through the inclusion of unconventional observations. The study was already discussed the context of temperature (see section 3.1), and the authors highlighted that by using observations of Q_{GW}, a significant number of unlikely flow regimes could be identified.

Applying automated inversion, Gannett et al. (2012) used a weighted multivariate objective function to simultaneously calibrate a regional IFM against observations of H and Q_{GW} to streams. Weights of the two different observation types were adjusted so that each observation type had a similar total weight in the multivariate objective function. Through this procedure, the successfully calibrated model could simultaneously reproduce climate- and pumping-induced fluctuations of both H and Q_{GW} on a multidecadal scale. The benefit of a weighted multivariate objective function that allows a model to be calibrated against all available observation types simultaneously was systematically demonstrated by Oehlmann et al. (2015), who investigated how the hydraulic parameters of karst systems can be better estimated with a combination of observations of H, Q_{GW} (spring discharge), and tracer breakthrough curves (i.e., TT information). They found that only a weighted multivariate objective function of H, Q_{GW}, and TT could properly calibrate the karst conduit network and hydraulic parameters and that sequential calibration with single-variable objective functions did not result in adequate models. Gagné et al. (2017) compared the manual calibration of a large, basin-scale GW flow model to automated calibration based on a weighted multivariate objective function, including both observations of H and Q_{base}. The performance of the two calibration approaches was assessed by comparing simulated to measured H, Q_{base}, and RT (derived from measurements of C). While the two models performed similarly in terms of the reproduction of H and Q_{base}, which were directly included in the calibration data set, they demonstrated that manual calibration performed significantly worse in the reproduction of observations of RT (which did not form part of the calibration data set). Automated calibration based on a weighted multivariate objective function reduced the mismatch between observed and

simulated RT by more than half compared to manual calibration. Gagné et al. (2017) demonstrated that GW RTs may be reproduced with a model calibrated against observations of H and Q_{base}, but only if that model is calibrated based on automated inversion of a weighted multivariate objective function—manual calibration failed in finding a parameter set to reproduce observations beyond the ones directly used in the calibration data set.

H. T. Li, et al. (2009) used a weighted multivariate objective function to automatically calibrate a regional IFM against observations of H and ET patterns obtained through remote sensing. They systematically assessed different ways of implementing ET patterns into the calibration process and used different combinations of H and ET. They found that if the depth to GW is smaller than the ET extinction depth, observations of H can be replaced by observations of ET wherever a reproduction of H is desired, no matter which type of ET pattern characterization they used. However, while observations of H could be accurately reproduced by using only ET observations in the calibration, simultaneously reproducing both H and ET was only possible using a combination of observations of H and ET. This study highlights the huge data worth of observations of ET for regional IFM calibration and the large potential to use observations of ET as complementary information alongside observations of H. That observations of ET are of high information content was also demonstrated by Schilling et al. (2014), who developed a relationship between tree ring growth, the depth to GW and plant transpiration for desert trees threatened by extreme droughts. Based on the established relationship, Schilling et al. (2014) quantified transpiration rates of the riparian forests along a reach of the Tarim River and calibrated a regional IFM against these observations alongside observations of H. Through a systematic assessment of different weighting strategies for the weighted multivariate objective function employed in the automated model calibration, the authors could show that observations of ET, even if they were associated with a very high uncertainty, were of pivotal importance in achieving an appropriate fit for both H and ET. The systematic data worth analysis was conducted using the linear predictive uncertainty analysis described in the tutorial of this review (see section 2). The data worth analysis revealed that with observations of H alone it was impossible to constrain the water balance and simultaneously calibrate K_{aq}, f, and soil and vegetation parameters. Only through inclusion of observations of ET alongside observations of H, the water balance could be sufficiently constrained, model parameters calibrated, and the predictive uncertainty of the model substantially reduced. Moreover, by successfully constraining the water balance, the predictive uncertainty of SW-GW exchange fluxes (infiltration of river water into the underlying aquifer) could also be reduced, and the water requirements of the desert trees estimated. The fact that observations of H are insufficient to calibrate parameters of unsaturated soils and that observations of ET allow closing the water balance and thus substantially reducing the ill-posedness of the inverse problem was also shown by Brunner et al. (2012) in a synthetic calibration experiment of a 1-D soil column. Using an automated calibration procedure with a weighted multivariate objective function and carrying out a systematic data worth and parameter identifiability analysis with the methods described by Doherty (2015), Brunner et al. (2012) showed that while observations of H allowed the model to be calibrated to reproduce observations of H, the parameter identifiability was weak and could only be improved through the inclusion of observations of ET. The inclusion of ET observations in the model calibration data set, however, did not substantially improve the reproduction of H in situations where depth to GW was high, which was also demonstrated by H. T. Li et al. (2009) who concluded that once the evaporation extinction depth is reached, observations of ET no longer contain any information about variations of H. Where the depth to GW was small, on the other hand, the information content of ET for the calibration of model parameters was substantially elevated in the studies of both H. T. Li et al. (2009) and Brunner et al. (2012).

3.2.3 Conclusions on Using Exchange Flux Observations

Unlike observations of T, which are confounded by many processes unrelated to the flow of water, observations of exchange fluxes are mainly controlled by hydraulic parameters, which make them ideal observation targets for flow model calibration. The studies that applied exchange flux observations clearly showed that the inclusion of these observations always improved the model calibration compared to using classical observations alone, despite the comparably large uncertainty associated with exchange flux measurements. Observations of Q_{GW}, Q_{base}, and ET were of particularly high data worth in the reviewed studies. However, accurate observations of exchange fluxes are more difficult to obtain than classical hydrological observations of H and Q_{SW}, and the increased uncertainty associated with observations of exchange fluxes must be taken into account during flow model calibration. Schilling et al. (2014) have demonstrated an efficient way to account for unknown uncertainty of exchange flux estimates by testing multiple weighting schemes in a postcalibration data worth analysis using the methods described in the tutorial of this review (see section 2).

The reviewed studies also showed that calibration against all observation types simultaneously using a weighted multivariate objective function resulted in the most appropriate model parameterizations, the most flexible way of integrating unconventional observations alongside classical ones and the most robust assessment of the benefit of using unconventional observations.

3.3 Soil Moisture

3.3.1 Observing Soil Moisture

Observations of H contain information about the saturated zone. Observations of soil moisture (S) provide complementary information for the unsaturated zone, which not only informs about SW-GW interactions, but also about subsurface-atmosphere-climate interactions (Seneviratne et al., 2010). Soil moisture observations can be obtained both at the point and at the regional scale, but in general, soil moisture observations are very uncertain (Lekshmi et al., 2014; Peng et al., 2017). Point measurements of soil moisture are typically influenced by small-scale processes, which are not normally reproduced by flow models (e.g., macropores and preferential flow paths), and therefore not suited for scaling to larger spatial scales (Hillel, 1998; Romano, 2014; Schilling, Irvine, et al., 2017). Remotely sensed observations of S are often used to infer spatially distributed locations and rates of R and Q_{GW} but are also notoriously uncertain (Brunner et al., 2004; Healy & Scanlon, 2010; Peng et al., 2017). The application of remotely sensed soil moisture observations was already discussed in the section on exchange fluxes (section 3.2) where soil moisture was used to infer R. But observations of S can also be directly used to calibrate flow models. In contrast to observations of R or Q_{GW}, however, to use observations of S directly, unsaturated flow processes need to be considered, which introduces additional parameters to the flow model.

3.3.2 Application of Soil Moisture Observations

Camporese et al. (2014) combined observations of H, Q_{SW}, and S to calibrate a regional IFM. Manual trial-and-error calibration of K and specific storage of two different zones (riparian and hillslope) was performed. This procedure allowed a parameterization that successfully reproduced all three observation types to be identified. The reduction in uncertainty or data worth of S was unfortunately not evaluated. Glaser et al. (2016) applied a manual trial-and-error procedure to calibrate K_{aq}, f, and various ET parameters of an IFM simulating a shallow, regional SW-GW system. Observations of Q_{SW} and two types of observations of S (i.e., soil moisture time series of the top 10 cm of soil and mapped surface saturation patterns) were used

sequentially during the calibration. The surface saturation patterns were integrated in the calibration by comparing simulated to observed saturated pixels in a visual output. The authors could show that the parameters that influenced the saturation patterns could be varied to match the patterns without substantially impairing the simulation of Q_{SW}. This illustrates that the information contained in observations of S can inform parameters that are relatively insensitive to Q_{SW} and H. In the synthetic study of Brunner et al. (2012) already discussed in the context of ET (section 3.2), remotely sensed observations of S (i.e., observations of S on the surface) were of less data worth compared to observations of H and ET. The worth of observations of both ET and S was higher for systems where the depth to GW was low. This highlights that the data worth of an observation type can vary with time.

3.3.3 Conclusion on Using Soil Moisture Observations

Among the unconventional observation types used in flow model calibration, observations of S are the least-frequently used. This could be due to the fact that if used directly, the observations of S require the representation of unsaturated flow processes in a flow model and the calibration of many additional parameters. However, the reviewed studies showed that, especially if remotely sensed, observations of S have a large data worth in regions where the depth to GW is low. A similar finding resulted from using remote observations of ET in flow model calibration studies. Observations of S thus provide complementary information about processes in the unsaturated zone to the information contained in classical observations. Unlike observations of ET, which can be obtained on similar spatial and temporal scales, observations of S do not provide information about fluxes. Observations of S nevertheless provide a valuable data source for flow model calibration.

3.4 Tracer Concentrations

3.4.1 Observing Tracer Concentrations

Observations of tracer concentrations (C) encompass a wide range of different types of tracers and observations, which can generally be made at a very high precision: They can, for example, pertain to situations where there is a known, local source of a tracer (e.g., a pollution plume, a tracer injection experiment, or perhaps a known area of stream infiltration that has an identifiable solute concentration), they can inform about a tracer source that is spatially invariable but temporally changing (e.g., tritium or a time marker such as low chloride concentrations due to an increase in recharge following clear cutting of a forested area), or the tracer could be produced in the subsurface in a spatially and temporally variable manner (e.g., ^{222}Rn, ^{37}Ar or ^{4}He; Cook & Herczeg, 2000; Schilling, Gerber, et al., 2017). Moreover, besides being available at a high precision and encompassing different kinds of tracers with different production, transformation, and transport pathways, the integration of observations of C into flow model calibration can be achieved in multiple ways: either directly through explicit mass transport calibration or indirectly without explicitly considering mass transport, for example, as derived quantities such as TT or RT information, as information on mixing of water from different sources or as information for the delineation of different water types. Depending on the tracer type and the desired type of observation, implementation of C into a flow model can be more or less straightforward: In the special case of a conservative tracer and where transport is predominantly driven by advection, flow tracking schemes that require significantly less parameters compared to explicit mass transport simulations might be justified. Unless such simplifications can be made, however, the simulation of tracers requires the explicit simulation of mass transport, which considers both advection and dispersion, and which requires additional model parameters to be defined. For observations of conservative tracers typically less parameters are required compared to observations of nonconservative

tracers (see Table 1). However, in some situations observations of C can be transformed prior to their implementation into the model calibration process, for example, to observations of exchange fluxes or observations of TTs or RTs. Such transformations reduce the number of processes and associated parameters that directly need to be integrated in the flow model. Overly simplifying assumptions during the transformation step, however, may lead to false model structures, which can be difficult to detect. Therefore, compared to classical observations and other unconventional observation types, prior to the integration of observations of C into flow model calibration, a careful evaluation of the underlying conceptual model and the different available integration procedures, including the potential pitfalls and inaccuracies associated with each procedure, is required.

Studies where observations of C were directly applied are reviewed in section 3.4.2, and studies that applied transformed observations of C such as flow path delineation or mixing observations are reviewed in section 3.4.3. While observations of TTs and RTs most often are transformed observations of C, they are associated with very particular transformation procedures and challenges during integration into flow model calibration and, for this reason, are discussed in a separate section devoted entirely to TTs and RTs (section 3.5).

3.4.2 Direct Application of Tracer Concentration Observations

In a synthetic flow and mass transport modeling study, Hendricks Franssen et al. (2003) quantified the data worth of direct observations of a conservative solute contamination plume in improving predictions of transport with the sequential self-calibration method (SSC; Hendricks Franssen et al., 2009). Observations of C were used alongside observations of H and information on K_{aq} to calibrate a transmissivity field. The best overall result was achieved by H and C in combination with the direct information on K_{aq}, whereas calibration against H alone and calibration against C alone resulted in poor parametrizations. As soon as observations of H or observations of C were complemented by direct information on K_{aq}, however, better parameterizations were obtained. Hendricks Franssen et al. (2003) showed that complementing observations of H with observations of C of a conservative solute substantially improved the predictive capability of the calibrated model but also that direct information of the respective parameters to be calibrated is of superior data worth to indirect observations (i.e., observations of system states such as H or C). In another synthetic flow and mass transport study, Xu and Gómez-Hernándenz (2016) employed EnKF to assimilate observations of a conservative solute contamination plume alongside observations of H and T in order to calibrate both K_{aq} and f. The combination of H, T, and C substantially reduced the uncertainty of K_{aq} and f. However, only using observations of C in the calibration data set resulted in the worst overall performance. Xu and Gómez-Hernándenz (2016) concluded that for the most realistic reproduction of heterogeneous structures in GW flow and mass transport problems, it is best to always assimilate H in combination with at least one unconventional type of observation. Alaghmand et al. (2014) and Alaghmand et al. (2016) did so and investigated the changes in floodplain salinity within a losing river-floodplain system in response to a lowering of the regional water table and to artificial river stage management. The manual trial-and-error calibration of an IFM against a combination of observations of H and C was successful in reproducing the propagation of I of river water into the floodplain and allowed recommendations for water and floodplain salinity management to be defined. Bonton et al. (2012) used an IFM to simulate nitrate transport and transformation in the capture zone of a drinking water well in an agricultural area. The authors showed that through the manual trial-and-error calibration of the IFM against not only observations of H but also against observations of C, changes in nitrate in the drinking water well as a function of agricultural use could be more accurately simulated compared to a calibration against only one of the two observation types. Mastrocicco et al. (2011) used a manual procedure to sequentially calibrate an IFM against observations of H, and

concentrations of Cl^- and NO_3^-. In contrast to Bonton et al. (2012), who also evaluated the benefit of NO_3^--based observations of C for manual model calibration, Mastrocicco et al. (2011) did not carry out a scenario analysis and calibrated the model manually and sequentially. Therefore, no statement on the worth of unconventional observations of C in improving their model could be made. On the same field site as Mastrocicco et al. (2011), Colombani et al. (2015) carried out a forced gradient pumping test and simulated the flow system with a heat and mass transport IFM. They calibrated their model manually and sequentially using observations of H and C of injected bromide. By including observations of C they could determine that only a dual-domain conceptualization of their flow system would reproduce the observed bromide concentrations and temperature patterns properly—this would not have been possible using only classical observations. Van der Hoven et al. (2008) attempted to calibrate a 2-D flow model of a river meander and underlying aquifer against observations of H, observations of C based on Cl^-, and observations of TTs. As with the studies of Mastrocicco et al. (2011) and Bonton et al. (2012), the purpose of the model was to investigate nitrogen cycling at an agriculturally impacted site. They conducted their calibration manually in multiple calibration scenarios, but according to the authors' own assessment, the procedure failed to reproduce observations of Cl^- sufficiently well. This was attributed to the lack of an appropriate heterogeneity of modelled K.

Hosseini et al. (2011) calibrated K, as well as dissolution and biodegradation rates of a nonconservative tracer plume of nonaqueous phase liquids (NAPLs), in a reactive mass transport model. In their synthetic study, Hosseini et al. (2011) first conditioned a heterogeneous K_{aq}-field to observations of H and direct information of K_{aq} using the SSC method and subsequently calibrated the dissolution and the biodegradation constants of NAPLs against observations of C. With this approach, the simulated reference concentration was successfully reproduced. Carniato et al. (2015) calibrated a reactive mass transport model against a combination of observations of H and C, but unlike Hosseini et al. (2011), Carniato et al. (2015) simulated a real-world case with real observations. They included direct information on K together with observations of H and of six different tracers to calibrate 369 hydraulic and chemical parameters using a weighted multivariate objective function with regularization. Using PEST, Carniato et al. (2015) conducted a comparison of posterior parameter uncertainty between a model calibrated only against observations of H, a model calibrated only against observations of C and a model calibrated against both observation types. They demonstrated that the uncertainty of the calibrated parameters was substantially smaller when both observation types were used simultaneously compared to only using one of them.

Boronina, Balderer, et al. (2005) manually calibrated a 2-D flow model against stable water isotope concentrations based on an explicit mass transport simulation. Compared to calibration of the flow model against H alone, which only allowed calibrating the ratio between transmissivities and R, only direct inclusion of C allowed recharge to be constrained and thus calibration of both transmissivity and recharge. The same model was extended into 3-D by Boronina, Renard, et al. (2005) and calibrated against observations of H, Q_{base}, Q_{GW}, and C in the form of ^3H measurements using a particle tracking scheme rather than explicit mass transport simulations. However, with their manual and sequential calibration the authors found that if K_{aq} was first calibrated against observations of H, Q_{base}, and Q_{GW}, subsequent calibration of f against observations of ^3H did not allow measurements of ^3H to be accurately reproduced—evidence that sequential calibration approaches can get trapped in local minima solutions, which only fit the first observation type used. Including observations of C simultaneously with the other observations could have helped avoiding an inadequate parametrization for ^3H transport. In a more successful sequential calibration, Gusyev et al. (2013) explicitly simulated transport of ^3H through a regional SW-GW system. They first calibrated K and R against

observations of H and Q_{base}, and subsequently *f* against observations of C. Gusyev et al. (2013) were able to reproduce the observed concentrations of ^{3}H. In yet another sequential calibration, Bauer et al. (2001) simulated GW flow plus nonreactive transport of ^{3}H, ^{85}Kr, chlorofluorocarbon (CFC)-113, and SF_6 with a 2-D watershed-scale IFM. They manually calibrated K against observations of H before manually calibrating *f* against the available tracer concentration observations. Simulated RTs were subsequently compared to measured RTs (based on ^{3}H). The authors could show that the calibration against so many different observation types allowed the reproduction of H and C, as well as adequate RTs. In a similar approach, Zuber et al. (2005) sequentially calibrated a flow and transport model against observations SF_6. Reproducing transport and TTs was not possible when the model was calibrated only against classical observations, and only the inclusion of observations of C improved the model parametrization. Unfortunately, missing information on the calibration procedure and on other data used to calibrate the IFM makes the results difficult to interpret. Mattle et al. (2001) used a reactive mass transport model of a real-world river-aquifer system to simulate the fraction of locally infiltrated river water in the pumps of an alluvial GW pumping station. In a first step, K_{aq} of two different aquifer materials and riverbed conductance were manually calibrated against observations of H. In a second step, riverbed conductance was refined manually so that the reactive transport simulations of tritiogenic ^{3}He concentrations of locally infiltrated river water matched respective observations. Mattle et al. (2001) found that riverbed conductance in alluvial SW-GW systems is largely insensitive to H of the floodplain and that only by including observations of C that inform locally infiltrating river water into the floodplain aquifer, riverbed conductance can be constrained so that SW-GW exchange fluxes are accurately simulated. Unfortunately, the manual calibration approach was not well documented. In a more recent, systematic and well-documented calibration study, Wood et al. (2017) calibrated a flow and transport IFM against observations of H and C obtained from ^{14}C measurements. They simultaneously calibrated K_{aq} and recharge of 18 different zones using a weighted multivariate objective function and an automated calibration procedure. Weights of the two observation types were set such that both observation types were accounted for equally (i.e., observation type balanced weighting). Based on the fact that the sensitivities of K_{aq} and R to concentrations of ^{14}C were large, the authors concluded that observations of C helped constraining the water balance of the investigated watershed beyond the capabilities of a model calibrated against observations of H alone. Castro et al. (1998) simulated GW flow and transport in the Paris Basin. In contrast to typical sequential calibration, where classical observations are used prior to unconventional observations, in their sequential and manual trial-and-error calibration routine Castro et al. (1998) first calibrated the IFM against ^{4}He concentrations and only in a second step verified the model with classical observations of H. K_{aq}, K of the aquitards, and the sources of ^{4}He could be successfully calibrated through this procedure and resulted in a good match between simulated and observed C and H. Castro and Goblet (2003) analyzed the nonuniqueness of a calibration of K for a regional IFM if only calibrated against observations of H and used concentrations of ^{4}He as an evaluation criterion. Calibration of K against observations of H alone resulted in many equally likely parametrizations but with very different distributions of K, and the subsequent calibration of transport parameters against observations of C revealed that only one of the identified parametrizations of K could reproduce the observations of C.

The study by Engelhardt et al. (2013), which was discussed in detail in the section on observations of T (section 3.1), is the only study that showed that observations of C contained less information than observations of T. However, the most likely reason for this is that their observations of C were based on concentrations of acesulfame, a tracer that was almost uniformly distributed throughout the studied system. The observations of C could therefore not inform about any flow or transport related processes beyond the capabilities of H and T, which themselves were characterized by much stronger spatial and temporal

variations.

3.4.3 Application of Transformed Observations of Tracer Concentrations

Rasa et al. (2013) investigated the data worth of observations of a conservative tracer in calibrating K and dispersivity of a flow and mass transport model. They simulated a long-term artificial tracer experiment and calibrated the IFM against observations of H and both direct and transformed observations of C using an automated calibration procedure based on a weighted multivariate objective function. The tracer measurements were used as three different types of observations: transient tracer concentration, absolute tracer mass discharge, and temporal first-order moments of the tracer breakthrough curves (i.e., TTs). In all three calibration experiments, one of the three observation types was paired with observations of H, and the three resulting calibrated models were compared. All three calibration experiments resulted in a significant reduction in predictive uncertainty of the model. Interestingly, using direct transient observations of C improved the model the least, whereas using transformed observations of C resulted in the more accurate reproductions of the observed tracer plume. Transformed observations of C thus contained more valuable information for the calibration of K and dispersivity in comparison to direct observations of C. In the flow, heat and mass transport modeling study already discussed in the context of T and exchange fluxes (sections 3.1 and 3.2), Delsmann et al. (2016) found that all tracer observations substantially improved the calibration of a SW-GW model, particularly in terms of the reduction of uncertainty of predictions of mass transport. They used both direct observations of C (i.e., electrical conductivity in drains and total salinity load in drains) and a transformed observation of C (i.e., depth of the salinity plume in the subsurface). The only observation types that could improve all model outcomes simultaneously were direct and transformed observations of C. Observations of H, T, and exchange fluxes worsened at least one prediction that was not related to the respective observation type. Like Rasa et al. (2013), Delsmann et al. (2016) found that if measurements of C were first transformed to salinity plume depth, calibration of the IFM was better than when direct observations of C were used. Hunt et al. (2006), in the study already discussed in section 3.2, also applied transformed observations of C for flow model calibration. Rather than explicitly simulating transport, they used an advective flow tracking scheme. Hunt et al. (2006) identified the depth of a lake water plume in the GW through the analysis of stable water isotopes, which allowed lake water to be differentiated from other sources of GW. The information about the lake water plume depth in the subsurface was then used as transformed observations of C. Despite the small number of plume depth observations compared to other observations (e.g., H), a systematic statistical analysis based on Cook's D showed that plume depths were among the most influential observations for the flow model calibration. Sanford et al. (2004) simulated a regional SW-GW system with the help of a flow tracking scheme, and in addition to H, calibrated their model against direct observations of ^{14}C activity as well as transformed observations of C based on the analysis of hydrochemistry. The hydrochemical information allowed different sources of GW to be distinguished, and this information was used to quantify the mix of water from different sources at each sampling location. The calibration was based on a multivariate objective function with error-based weights. By combining observations of H, direct observations of C, and transformed observations of C, the model could be calibrated so that both spatially distributed recharge and the mix of water could be reproduced. Doyle et al. (2015) used observations of H in combination with observations of recharge locations derived from noble gas concentrations to manually calibrate a 3-D model with particle tracking. The inclusion of recharge locations as calibration targets significantly helped constrain K and recharge locations, especially in regions where no observations of H were available. The studies of Sanford et al. (2004) and Doyle et al. (2015) demonstrate that a well-founded transformation of observations of C to source delineation or mixing information enables the use of a computationally much less intensive transport representation via particle tracking, rather than the

computationally intensive explicit transport simulations often required for direct observations of C. The same approach was used by Schilling, Gerber, et al. (2017), who, in a first step, calibrated a 3-D real-world IFM of a prealpine river-aquifer system with a flow tracking scheme against observations of H using an automated calibration procedure and, in a second step, systematically compared the resulting parametrizations in their ability to reproduce transformed observations of C (i.e., SW-GW mixing ratios and RT). Like in the study of Mattle et al. (2001) that was discussed in section 3.4.2, one of the main goals of the study of Schilling, Gerber, et al. (2017) was to quantify the amount of locally infiltrating river water in the pumped drinking water of an alluvial GW pumping station. While the automated calibration procedure allowed K_{aq}, K_{rb}, and f to be calibrated to reproduce observations of H, subsequent comparison to transformed observations of C in the form of (1) mixing ratios between infiltrated river water and regional GW and (2) RTs of infiltrated river water showed that only by considering transformed observations of C, K_{rb} and f could be sufficiently constrained to not only reproduce H but also SW-GW exchange fluxes, mixing ratios, and RTs. The simulation of H was largely insensitive to K_{rb} and f and observations of H were thus unable to constrain K_{rb} and f sufficiently well for the reproduction of SW-GW interactions in an alluvial system. The systematic analysis revealed that K_{rb} could be constrained by observations of the fraction of recently infiltrated river water in GW, whereas f was sensitive to observations of RT of locally infiltrated river water. These findings are very similar to the findings of Mattle et al. (2001) who also found that K_{rb} was largely insensitive to H but the key parameter for SW-GW exchange fluxes within an alluvial valley, and uniquely identifiable only with the help of observations of tracer concentrations.

3.4.4 Conclusions on Using Observations of Tracer Concentrations

The above findings show that the calibration of IFMs against observations of C harbors huge potential. In almost every study, the inclusion of observations of C substantially improved the parametrization of IFMs. Moreover, the studies that systematically compared data worth revealed that rather than just improving one single type of prediction, tracer concentrations are the only observation type able to improve all types of model predictions simultaneously. The benefit of including direct observations of C in IFM calibration appears to outweigh the cost of requiring additional parameters for explicit mass transport simulations. In the special situations where observations of C could be transformed to other observations, for example, to information on the sources of GW, this not only allowed the computational demands to be reduced by eliminating the need of explicit mass transport simulations but also resulted in improved data worth. In the studies where the data worth of direct observations of C was systematically evaluated and compared to the data worth of transformed observations of C, the transformed observations were of even larger data worth. As with observations of exchange fluxes, observations of C, whether they are used directly or whether they are transformed prior to their implementation into flow model calibration, were an extremely beneficial type of unconventional observation.

The above studies also shed light on another important aspect of flow model calibration: The studies of Boronina, Renard, et al. (2005) and Castro and Goblet (2003) not only highlight the nonuniqueness of a calibration of K based on purely classical observations—they also demonstrate the potential pitfalls of sequential model calibration. If, for example, K is calibrated against classical observations in a first step and then fixed during subsequent calibration steps, the potential of choosing an inappropriate parameterization is large, and the potential of finding a better solution for K with unconventional observations is eliminated. If unconventional observations are used alongside classical observations for flow model calibration, the calibration routine should be based on (regularized) weighted multivariate objective functions that allow the simultaneous calibration of all model parameters against all observation types. The benefit of this procedure

in relation to observations of C was, for example, demonstrated by Sanford et al. (2004), Hunt et al. (2006), Rasa et al. (2013), Carniato et al. (2015), and Wood et al. (2017). In their guideline, Doherty and Hunt (2010) provided specific suggestions for the weighting of observations of C if used alongside observations of H: They suggest (1) ensuring that the total weight of all observations of C is similar to the total weight of all observations of H so that both groups are equally important in the calibration objective function and (2) setting the weights among contamination plume concentration observations so that small concentrations marking the outlines of the contamination plume are more strongly visible in the calibration objective function, as these observations inform most about local subsurface heterogeneity.

3.5 Travel Times and Residence Times

3.5.1 Observing Travel Times and Residence Times

The TT of a water parcel in the subsurface refers to the time that GW takes to travel between two discrete locations within the subsurface. The RT, or "groundwater age," of a water parcel refers to the total time it takes for GW to travel between the recharge or infiltration point to the sampling point. As aquifers are typically characterized by spatially distributed recharge rather than recharge from a single point, are highly heterogeneous, and as GW within an aquifer is subject to mixing processes of different temporal and spatial scales, GW samples always represent a mix of water with different RTs. A RT estimation of a sample made with a single tracer therefore reflects the mean of a distribution of RTs on the scale to which that tracer is sensitive and is thus referred to as "apparent groundwater age." The fact that a GW sample always represents a mix of water with different RTs, and because all tracers have a specific range of RTs to which they are sensitive, multiple tracers are often combined into a multitracer-based RT analysis, which allows a better estimation of the RT distribution within a GW sample (e.g., Bauer et al., 2001; Schilling, Gerber, et al., 2017). A detailed discussion on measuring RTs is beyond the scope of this review, and for more detailed discussions on the nature, application, and interpretation of RT measurements, see Cook and Herczeg (2000), McCallum et al. (2014), McCallum et al. (2015), Purtschert (2008), or Sanford (2011).

TTs are typically obtained from artificial tracer injection experiments, in which a known quantity of a conservative solute is injected at one location and measured at discrete downstream sampling locations. RTs are typically estimated through the measurement of environmental tracers that are already present in the environment and that allow the back-calculation of the time when the water was last in contact with the atmosphere (e.g., based on ^{222}Rn, ^{37}Ar, ^{3}H/^{3}He, ^{14}C, ^{3}H, and CFC). To obtain observations of TT and of RT, measured concentrations are transformed based on comparably simple mathematical and physical models, for example, using simple exponential production and decay laws under the assumption of homogeneity and of full saturation of the aquifer. This transformation step from observations of C to TTs and RTs introduces additional uncertainty to TT and RT observations.

Like direct observations of C, the integration of observations of TT between two discrete points within a GW flow system into the calibration of an IFM is relatively straightforward and under favorable conditions (i.e., under predominantly advective transport) can be done with an advective flow tracking scheme rather than with explicit advection-dispersion-based mass transport and RT simulations. Compared to the direct use of observations of C, however, using observations of RT (and to a lesser extent of TT) for IFM calibration is less straightforward and harbors the danger of introducing a bias toward younger RT, as most natural environmental tracers are limited in the coverage of RTs (Gardner et al., 2013; McCallum et al., 2014, 2015; Sanford, 2011): Once the maximum estimable RT of a tracer is reached by at least some fraction of the sampled water, the additional time spent in the subsurface is not identifiable by that tracer anymore, and the

estimated apparent GW age thus often appears too young. For systems where high GW ages are expected and where the applied tracers do not cover all expected RT, directly simulating the (reactive) transport of the tracer used for calibration is recommended rather than using simpler flow and particle tracking approaches (McCallum et al., 2015). A way of minimizing this bias can be achieved by applying a multitracer approach so that the applied tracers (1) cover all relevant RTs, (2) water types can be differentiated, and (3) water-type specific RTs can be obtained (see Schilling, Gerber, et al., 2017). Sanford (2011), moreover, pointed out that if calibration is based on automated inversion using a weighted multivariate objective function, the weight that is given to observations of TT and RT can be chosen so that it represents the uncertainty and potential pitfalls that are associated with observations of TT and RT, thus avoiding overfitting and an introduction of bias.

3.5.2 Application of Travel Time Observations

As discussed in the previous section on observations of C (section 3.4), Rasa et al. (2013) showed that in terms of reducing the predictive uncertainty of a GW model that could be extracted from an artificial tracer test, the most informative observation type was the first-order moment of the tracer concentration breakthrough curve, that is, the mean TT. The calibration of K and dispersivity were improved when the multivariate objective function of the flow and transport model included observations of H and TT, compared to a combination of H and C. Similarly, Oehlmann et al. (2015) used peak arrival times of tracer breakthrough curves as observations of TT for calibration of a karst system. In their study, which was already discussed in section 3.2, calibration against a combination of observations of H and Q_{GW} with TT allowed structural errors in the model that were not visible solely by comparing simulated versus observed H to be identified. Only a change of the initial model structure allowed all three components (H, Q_{GW}, and TT) of the multivariate objective function to be simultaneously satisfied. The information contained in observations of TT in finding an appropriate model structure and parametrization was of pivotal importance for the studied karst system. Hunt et al. (2006) estimated the TT of infiltrating lake water to a set of wells based on measurements of ^{3}H and CFCs. The estimated TT was subsequently included in the calibration of a regional SW-GW flow model with flow tracking alongside many other observation types (discussed in the sections 3.2 and 3.4). However, due to strong assumptions that are required to interpret TT or RT, Hunt et al. (2006) only applied a comparably low weight to the TT observation in their calibration objective function (as recommended by Sanford (2011)). As a result, the observation of TT did not significantly contribute to the calibration of any parameter. Thus, the information content of an observation type is dependent on the flow system in question, the assumptions and resulting uncertainties associated with each observation, and the weight that is eventually given to each observation during the flow model calibration.

3.5.3 Application of Residence Times Observations

In probably the earliest study in which unconventional observations were used to calibrate a complex GW model, Reilly et al. (1994) manually calibrated a flow and mass transport model against observations of H and RT based on measurements of CFCs and ^{3}H. An initial calibration against H resulted in a parametrization that was not able to reproduce RT. The model was recalibrated against observations of H and RT simultaneously, which resulted in a parameterization that successfully reproduced not only H and RT but also the measured concentrations of ^{3}H. Not long thereafter, Sanford and Buapeng (1996) used RT inferred from ^{14}C measurements to estimate paleo-flow conditions, which they then used to define the most appropriate structure of a regional GW flow model. In contrast to model structures that were not informed by the RT estimates, the model structure that accounted for paleo-flow conditions was the only one to successfully reproduce mean RT in a flow and flow tracking framework. Observations of RT were thus successfully used

to inform the conceptual model and the model design, rather than directly used in the calibration procedure. Zhu (2000) also calibrated a flow model against observations of H and of RT based on ^{14}C measurements to define appropriate flow conditions. The flow model was calibrated manually by visually comparing model outputs to measurements. While this procedure allowed a suitable set of hydraulic conductivities and recharge rates to be defined, as with all manual calibration, this procedure is not rigorous enough to allow for quantitative conclusions about the information content of observations of RT. In a more rigorous study, Sanford et al. (2001) and Sanford et al. (2002) calibrated a flow model against observations of H and RT based on measurements of ^{14}C and an automated calibration routine. This two-part study is one of the earliest studies in which a large number of parameters was simultaneously calibrated against observations of H and RT using a weighted multivariate objective function. The authors concluded that including observations of RT constrained the water balance and allowed absolute parameter values rather than parameter ratios to be identified, which would have been the case if only observations of H were used. Michael and Voss (2009) attempted to manually calibrate a large regional GW flow and flow tracking model to observations of H and RT derived from ^{14}C measurements. A sensitivity analysis showed that the first attempt using only observations of H in shallow wells allowed just three out of the necessary 32 parameters to be calibrated. In a second attempt, the model was manually calibrated against observations of RT, which resulted in a much better reproduction of both H and RT. Michael and Voss (2009) pointed out that the simplifying assumptions that needed to be applied to the RT estimation from tracer measurements, as well as to the RT model simulation potentially resulted in substantial structural inaccuracies. Nevertheless, due to the high sensitivity of K to observations of RT and due to the similarity in the optimized parameter values, the authors concluded that observations of RT should be included more frequently in flow model calibration. In the multiobservation study by Manning and Solomon (2005) (discussed in sections 3.1 and 3.2), the inclusion of observations of RT into the manual calibration of a flow model was key in distinguishing between inappropriate and appropriate model parametrizations for a regional flow and transport model of the Salt Lake Valley, Utah. In the study by Schilling, Gerber, et al. (2017), which was already discussed in the context of transformed observations of C (see section 3.4.3), only the consideration of observations of RT of locally infiltrated river water obtained in a multitracer approach that covered all relevant RTs allowed f to be constrained alongside K_{aq} and K_{rb}.

3.5.4 Conclusions on Using Observations of Travel Times and Residence Times

The above findings confirmed that observations of TT and RT can be beneficial for flow model calibration but that their implementation into a flow model calibration routine requires a significant number of assumptions to be made and, therefore, is associated with uncertainties. Using RT and TT observations can be achieved with a simplified flow tracking routine rather than requiring computationally much more demanding explicit mass transport simulations, as it is required for the calibration of a flow model against observations of tracer concentrations. However, compared to explicit mass transport simulations, if simplified flow tracking routines are used the measured tracer concentrations first need to be converted to an RT or TT observation according to predefined mathematical and conceptual models, which are never perfect (see McCallum et al., 2014, 2015). For this reason, Sanford (2011) reminded modelers to take care to not "overfit" flow models to RT observations. Despite these potential pitfalls, the reviewed studies showed that observations of RT and TT, if used conservatively, help significantly in eliminating inappropriate model parametrizations, which would otherwise not be detectable.

Only Sanford et al. (2001, 2002) conducted calibration of a flow model against observations of RT based on automated calibration routines. All other models reviewed in the context of observations of RT were

calibrated manually, which, as outlined in the introduction, does not allow for the optimal set of parameters to be found or for a quantitative analysis of the information content of RT observations. Gagné et al. (2017) (discussed in section 3.2) used observations of RT to systematically compare the outcome of a manual flow model calibration against an automated flow model calibration, but unfortunately observations of RT were only used in the validation step and the calibration was based on observations of H and Q_{GW}. In order to allow for more concise conclusions about the use and information content of observations of RT for flow model calibration, future studies should include observations of RT in the calibration data set and be based on automated calibration routines.

4 Summary and Discussion

This review fills a gap in the discussion of the broad and diverse topic of flow model calibration. In general, the reviewed studies confirm that including at least one unconventional observation type of a relevant process for the modeled system (and for the desired type of prediction) alongside classical observations strongly reduces the ill-posedness of the inverse problem, that is, improves parameter identifiability and reduces the uncertainty of predictions made with the flow model. Reducing the ill-posedness of flow models by inclusion of additional and unconventional observations into the calibration process is thus the preferential alternative to artificially reducing the ill-posedness of GW inverse problems. Artificially reducing model complexity by fixing or homogenizing parameters of flow models strongly limits the ability to explore the influence of heterogeneity on flow model predictions, whereas adding observations to the calibration data set reduces the model predictive uncertainty under consideration of complexity and heterogeneity (e.g., Gianni et al., 2018).

In terms of the data worth of the different unconventional observations in reducing the uncertainty of predictions of fluxes and tracer concentrations, the findings summarized in this review strongly suggest that observations of exchange fluxes and tracer concentrations contain the largest amount of complementary information to classical observations. Also, observations of RTs and TTs contain valuable information about SW-GW flow systems, but their implementation into flow model calibration needs to be done with great care, as it is associated with many potential pitfalls. Unlike exchange flux observations, which do not rely on many underlying assumptions and do not require additional processes to be simulated, the successful implementation of observations of RT and TT into flow model calibration is typically based on many underlying assumptions on the conceptual model and on simplified representations in flow models. Observations of temperature, on the other hand, are relatively simple to implement but often confounded by processes that are not related to the transport of water in order for them to be beneficial for flow model calibration: more often than not, calibration against observations of temperature resulted in improvements of predictions of heat flux only, while the quality of most other predictions worsened. Temperature observations were beneficial for flux and transport calibration only if they related to exchange fluxes between SW and GW bodies, and only if taken within 2m of the interface between the two water bodies.

The overall predictive uncertainty of a model was often reduced the most through unconventional observation types that were transformed prior to their inclusion into the calibration procedure. In one example, using salinity measurements to delineate interfaces between freshwater and saltwater in the case of saltwater intrusion and then using the delineation as an observation more strongly reduced the predictive uncertainty of the model than directly using salinity in the objective function (see Delsmann et al., 2016). Or instead of using observations of noble gas concentrations directly, using observations of recharge elevations that can be derived from noble gas concentrations in the objective function can facilitate flow model calibration (see Doyle et al., 2015). However, one should exercise due care when choosing to apply a transformation or a

simplified process representation in flow models: Depending on the spatial and temporal scale and the nature of the original observation, transformations and simplifications might result in an oversimplification, and calibration might result in a biased model, of which the structural defects cannot be quantified (see Turnadge & Smerdon, 2014). Transformed observations should thus only be considered for cases where direct integration of the untransformed observations is impossible or unfeasible, or where the transformation is not associated with substantial additional uncertainty. Nevertheless, transformed observations can substantially improve the application of unconventional observation types in flow model calibration. This is particularly evident if the transformation results in observations that do not require additional processes such as heat or mass transport to be simulated and can be represented in a simplified manner, for example, with a flow tracking scheme.

The above classification of the worth of different observation types is of qualitative nature due to the fact that the information content of different observation types and individual observations varies with (1) the system under investigation, (2) the temporal and spatial scales of interest, (3) the modeling purpose (i.e., the desired predictions), and (4) the modeling and calibration strategy. The three challenges of including unconventional observations into flow model calibration outlined in the introduction further influence the worth of different observation types and individual observations. It is thus impossible to derive a quantitative metric of the information content of each observation type and even more so of an individual observation that would be generally valid for every modeling study. While not resulting in one single, generally valid quantitative metric, the present review allowed the identification of particularly versatile and information-rich observation types and suitable strategies for implementing unconventional observations. Weighted multivariate objective functions that allow simultaneous calibration against diverse observation types using automated mathematical calibration routines could be identified as highly suitable for the calibration of flow models. Both frequentist and Bayesian methods are suitable mathematical algorithms to solve the inverse problem and for predictive uncertainty analyses. Manual trial-and-error calibration can provide a good estimate of the importance of certain observations and parameters, but the reviewed studies clearly show that the technique is not suited for complex flow models or large data sets and does not allow a mathematically robust quantification of optimal parameter sets, predictive uncertainty, or data worth. Manual trial-and-error calibration should only be used as a preliminary tool to assess optimal calibration and observation implementation strategies. For data worth analyses, true Bayesian methods are more robust but computationally more demanding so that, until now, they have been generally only applied for data worth analyses in synthetic flow modeling studies. Frequentist methods that require the calculation of sensitivities between model outputs and parameters provide efficient means to calculate data worth. However, the efficiency comes at the cost of being less robust compared to true Bayesian methods. Besides following the qualitative ranking provided above, the ideal observation type for a given flow model study should directly inform the process of interest; that is, if heat transport is a prediction of interest, observations of temperature are likely the most informative unconventional observation type. For hydraulic head, water (exchange) flux, and mass transport predictions, observations of exchange fluxes and tracer concentrations (both direct observations and transformed observations) are most beneficial.

Since only a few of the reviewed studies carried out a systematic analysis of uncertainty and data worth, there remains a need to better characterize ideal unconventional observation types and implementation methods. To the best of our knowledge, a systematic quantification of the worth of different observation types in calibrating a flow model was only done by Brunner et al. (2012), Delsmann et al. (2016), Engelhardt et al. (2013), Hendricks Franssen et al. (2008), Hendricks Franssen et al. (2003), Hunt et al. (2006), La Vigna et al. (2016), Masbruch et al. (2014), Rasa et al. (2013), Carniato et al. (2015), and Schilling et al. (2014). A clear need for systematic data worth and uncertainty analyses can thus be identified, particularly when

considering that models can never be regarded as a perfect representation of reality. Predictive uncertainty is of pivotal importance for decision makers, and information on the worth of different observation types is essential in order to optimally plan data acquisition campaigns (Brunner et al., 2012; Dausman et al., 2010; Doherty, 2015; Doherty & Welter, 2010; Hill & Tiedeman, 2007; White et al., 2016). The current generation of calibration tools, including PEST (Doherty, 2015) and UCODE (Poeter et al., 2014), provides the means to analyze data worth in a systematic way using the calculated sensitivities between model parameters, model outputs, and observations of the automated calibration with little postprocessing efforts.

Future research should be directed toward an improved assessment of (1) the best way of model-data integration and (2) the data worth of unconventional observation types for different flow systems so that modelers can better decide, prior to model development, which observations to obtain for the system of interest and goal of the model, and on how to best include these unconventional observation types into the flow modeling process. If this will be achieved, the systematic application of unconventional observations in flow model calibration is going to hugely improve the quality of simulations and predictions of SW-GW flow systems.

Notation

Physical variables

K
: hydraulic conductivity (LT^{-1})

K_{aq}
: aquifer hydraulic conductivity

K_{rb}
: riverbed hydraulic conductivity

S
: storage parameter (−; storativity for confined, specific yield for unconfined systems)

f
: effective porosity (−)

α_{sol}
: (longitudinal/transverse) dispersivity in porous medium (L)

$\alpha_{sol,SW}$
: (longitudinal/transverse) dispersivity in surface water (L)

D_{sol}
: diffusion in the porous medium [L^2T^{-1}]

$D_{sol,SW}$
: diffusion in surface water [L^2T^{-1}]

λ_{source}
: mass production coefficient in the porous medium [T^{-1}]

$\lambda_{source,SW}$
: mass production coefficient in surface water [T^{-1}]

λ_{sink}
: decay/adsorption coefficient in the porous medium [T^{-1}]

$\lambda_{sink,SW}$
: decay/adsorption coefficient in surface water [T^{-1}]

α_{heat}
: thermal dispersivity of the porous medium [L]

$\alpha_{heat,\ SW}$
: thermal dispersivity in surface water [L]

κ_{bulk}
: bulk thermal conductivity [L^2T^{-1}]

ρ_{bulk}
: density of the bulk material [ML^{-3}]

c_{bulk}
: specific heat capacity of the bulk material [$L^2M^{-2}\Theta^{-1}$]

n
: friction coefficients (e.g., Manning roughness coefficients [$L^{-1/3}T$])

h_d
: rill/depression storage height

h_o
: obstruction storage height

Boundary conditions

H_{BC}
: fixed hydraulic head boundary condition (i.e., first order boundary condition) [L]

Q_{BC}
: fixed flux boundary condition (i.e., second order boundary condition) [L^3T^{-1}]

C_{BC}
: fixed concentration boundary condition [ML^{-3}]

T_{BC}
: fixed temperature boundary condition [Θ]

Abbreviations

DTS
: distributed temperature sensing

C
: tracer concentration

EnKF
: ensemble Kalman filter

FO
: fiber optics

EC
: electrical conductivity

ET
: evapotranspiration

GW
: groundwater

H
: hydraulic head

I
: infiltration

IFM
: integrated flow model

MLE

maximum likelihood estimation

NAPL
: nonaqueous phase liquids

NGRT
: noble gas recharge temperature

Q_{base}
: baseflow

Q_{GW}
: groundwater discharge/springs

R
: recharge

RT
: residence time

S
: soil moisture

SSC
: sequential self-calibration

SW
: surface water

T
: temperature

TT
: travel time

Acknowledgments

We want to thank John Doherty, Harrie-Jan Hendricks Franssen, and Daniel Hunkeler for their valuable inputs and comments. We would also like to thank the editors Fabio Florindo and Gregory Okin; the associate editors; and Michael Cardiff, Randy Hunt, Andrea Brookfield, and three anonymous reviewers for their positive feedback and insightful comments. Oliver S. Schilling gratefully acknowledges the funding provided by the Swiss National Science Foundation grant P2NEP2_171985. Only published data were used to produce this review.

References

Citing Literature

Download PDF

Journal | Articles Actions

Back to Top