AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016476

Key Points:
- We evaluated the value of additional data to constrain a groundwater model
- Disinformation in observational data obscured clear value of additional data types
- Diverse conditioning data types were essential to constrain transport modeling

Supporting Information:
- Supporting Information s1

Correspondence to:
J. R. Delsman,
joost.delsman@deltares.nl

Citation:
Delsman, J. R., P. Winters, A. Vandenbohede, G. H. P. Oude Essink, and L. Lebbe (2016), Global sampling to assess the value of diverse observations in conditioning a real-world groundwater flow and transport model, *Water Resour. Res.*, *52*, 1652–1672, doi:10.1002/2014WR016476.

Received 29 SEP 2014
Accepted 31 JAN 2016
Accepted article online 5 FEB 2016
Published online 6 MAR 2016

Global sampling to assess the value of diverse observations in conditioning a real-world groundwater flow and transport model

Joost R. Delsman[1,2], Pieter Winters[1,3], Alexander Vandenbohede[4,5], Gualbert H. P. Oude Essink[1,6], and Luc Lebbe[7]

[1]Unit Subsurface and Groundwater Systems, Deltares, Utrecht, Netherlands, [2]Critical Zone Hydrology Group, Department of Earth Sciences, VU University Amsterdam, Amsterdam, Netherlands, [3]Now at DEME Group, Zwijndrecht, Belgium, [4]Department Geology and Soil Sciences, Ghent University, Gent, Belgium, [5]Now at De Watergroep, Brussels, Belgium, [6]Department of Physical Geography, Utrecht University, Utrecht, Netherlands, [7]Groundwater Modelling Cell, Department Geology and Soil Sciences, Ghent University, Ghent, Belgium

Abstract The use of additional types of observational data has often been suggested to alleviate the ill-posedness inherent to parameter estimation of groundwater models and constrain model uncertainty. Disinformation in observational data caused by errors in either the observations or the chosen model structure may, however, confound the value of adding observational data in model conditioning. This paper uses the global generalized likelihood uncertainty estimation methodology to investigate the value of different observational data types (heads, fluxes, salinity, and temperature) in conditioning a groundwater flow and transport model of an extensively monitored field site in the Netherlands. We compared model conditioning using the real observations to a synthetic model experiment, to demonstrate the possible influence of disinformation in observational data in model conditioning. Results showed that the value of different conditioning targets was less evident when conditioning to real measurements than in a measurement error-only synthetic model experiment. While in the synthetic experiment, all conditioning targets clearly improved model outcomes, minor improvements or even worsening of model outcomes was observed for the real measurements. This result was caused by errors in both the model structure and the observations, resulting in disinformation in the observational data. The observed impact of disinformation in the observational data reiterates the necessity of thorough data validation and the need for accounting for both model structural and observational errors in model conditioning. It further suggests caution when translating results of synthetic modeling examples to real-world applications. Still, applying diverse conditioning data types was found to be essential to constrain uncertainty, especially in the transport of solutes in the model.

1. Introduction

Mathematical modeling of the complex and dynamic processes that drive the transport of groundwater and associated solutes to surface water necessarily involves simplification and hence uncertainty. And while some techniques can provide proxy information on subsurface structure (e.g., airborne geophysics [*Gunnink et al.*, 2012] and seismic tomography [*Cassiani et al.*, 1998]), properties of the (sub)surface are generally unmeasurable at the desired model scale, let alone the effective model parameters that describe these properties [*Beven*, 1989]. Estimation of model parameters using available observational data is therefore common practice in groundwater modeling, be it through manual (trial-and-error) calibration or mathematically solving the inverse problem [*Carrera et al.*, 2005; *Zhou et al.*, 2014].

Especially in distributed groundwater models, limited observational data and correlation between the large number of model parameters often leave the inverse problem ill posed. Reduction of model parameters has been sought in model parsimony [*Hill*, 2006], zonation of parameters [*Hill and Tiedeman*, 2007], or through parameter regularization [*Tonkin and Doherty*, 2005; *Hunt et al.*, 2007; *Doherty et al.*, 2011]. On the other hand, many researchers have suggested using additional observational data to constrain the inverse problem. Among other concentration data [e.g., *Barlebo et al.*, 1998, 2004], conservative tracers [e.g., *Anderman et al.*, 1996; *Barth et al.*, 2001; *Rasa et al.*, 2013], temperature [e.g., *Bravo et al.*, 2002; *Risley et al.*, 2010], geophysical data [*Beaujean et al.*, 2014], age tracers [e.g., *Ginn et al.*, 2009; *Gusyev et al.*, 2013; *Nassar and Ginn*,

2014a,b], gravity data [*Christiansen et al*., 2011; *Sun et al*., 2012], and combinations thereof [*Hunt et al*., 2006; *Vandenbohede et al*., 2011] have all been applied alongside commonly used head measurements in groundwater model calibration. In this paper, we investigate the added value of discharge data, concentration data, geophysical measurements, and temperature data in calibrating, or conditioning, a groundwater flow and transport model.

Automatic parameter estimation in groundwater modeling is generally performed using weighted nonlinear regression, implemented in codes such as UCODE [*Poeter and Hill*, 1999] or PEST [*Doherty*, 2010]. Weighted nonlinear regression estimates model parameters conditional on a hypothesized model structure and a priori characterized observational uncertainty, by seeking an optimum in some defined objective function. The method, its assumptions, and good practices are extensively treated in *Hill and Tiedeman* [2007]. *Beven* [2006] observed that equally acceptable performing models can be generally found from many different regions of the model parameter space, contradicting the idea in nonlinear regression of an "optimal model." The generalized likelihood uncertainty estimation (GLUE) methodology [*Beven and Binley*, 1992] is based on this notion of "equifinality." GLUE uses Monte Carlo sampling to globally search the model (parameter) space and rejects models that are not "behavioral," i.e., that do not correspond well enough to observations given a, subjective, a priori defined rejection criterion. Predictions are based on the entire set of remaining behavioral models, weighed according to some (often informal) likelihood measure [*Beven and Binley*, 1992; *Beven*, 2006, 2009]. Despite having received criticism for the subjective, nonformal statistical treatment of uncertainty (most notably by *Mantovan and Todini* [2006], *Stedinger et al*. [2008], and *Clark et al*. [2011]), GLUE has found widespread use [*Beven and Binley*, 2013]. However, its computational demands have limited reported groundwater examples to only a few [e.g., *Feyen et al*., 2001; *Hassan et al*., 2008; *Rojas et al*., 2008].

Several studies have quantified the value of different observational data in constraining parameter uncertainty in groundwater models. With nonlinear regression being the dominant parameter estimation method in groundwater modeling, most studies used regression results and associated statistics to report on data value [e.g., *Barlebo et al*., 1998; *Barth and Hill*, 2005b; *Hunt et al*., 2006]. They commonly found that adding different observational data to the calibration process improved the simulation of the modeled system, was necessary to calibrate modeled processes not captured by conventional targets, better constrained parameter estimates, narrowed confidence intervals and facilitated parameter convergence. These results are, however, conditional on a found parameter optimum and usually involve linearization of the response surface of the objective function around this optimum. Results therefore do not necessarily hold for possible equally fit models in different regions of the parameter space, as the shape of the response surface will likely vary [*Beven*, 2009; *Rakovec et al*., 2014]. The use of more computationally demanding parameter estimation methods, to more inclusively account for model and measurement uncertainty, is increasingly reported in the literature [e.g., *Feyen et al*., 2003; *Hendricks Franssen et al*., 2003; *Keating et al*., 2010; *Rojas et al*., 2010; *Rasa et al*., 2013; *Carniato et al*., 2014]. However, studies that compare different observational data types using more computationally demanding parameter estimation methods are generally based on synthetic model experiments that necessarily idealize the complex reality and disregard disinformation in the observational data. Results may therefore not be straightforwardly applied to real problems [*Beven et al*., 2008].

1.1. Scope and Outline

In this paper, we aim to (1) establish the value of different data types and combinations thereof in conditioning a groundwater flow and transport model of an agricultural field site in the Netherlands [*Delsman et al*., 2014a], and (2) investigate the possible confounding effects of disinformation in the observational data on model conditioning. We did not further investigate causes of disinformation (be it errors in model structure, for instance, due to geologic uncertainty, or errors in the observational data), this was considered outside the scope of the present paper. Nor did we directly compare GLUE and more computationally frugal methods.

The data available for the extensively monitored site allowed us to investigate the value of a wide variety of data types (including heads, flow, concentrations, and temperature) in model conditioning. We structured our analysis (and this paper) as follows: we first constructed a groundwater flow and transport model for our field site. Second, we used the GLUE methodology to assess the value of different observational data

types and combinations thereof in conditioning our model first in a synthetic modeling experiment. Third, we performed the same analysis but now using the actual observational data. Comparing results from this real-world case to its synthetic variant allowed us to investigate the influence of disinformation in observational data in model conditioning. In a final step, we combined data types to best condition our model to the studied agricultural field site, and compared results to a heads-only-conditioned model.

2. Methods

2.1. Study Area and Measurement Setup

We studied a 35 m slice of a 900 m × 125 m agricultural field, located 20 km north of Amsterdam, Netherlands (52.599° lat, 4.782° lon). A full description of the field site, measurement setup, and measurement results has been presented elsewhere [*Delsman et al*., 2014a]. For brevity, only a brief summary is repeated here. The field site is situated in the Schermer polder, a former lake reclaimed in 1635 A.D. (Geo)hydrological conditions are typical for polders in the coastal region of the Netherlands. The average annual precipitation surplus of 290 mm (precipitation 880 mm, evapotranspiration 590 mm) is drained by a system of tile drains (every 5 m, 1 m depth) and ditches, limiting groundwater level variation to within 0.6 and 1.6 m below ground surface (BGS) [*Delsman et al*., 2014a]. During the growing season (April–September), extraneous fresh river water is diverted into these ditches to maintain water levels and dilute ditch salinities to enable sprinkling irrigation. Near-surface geology of the field is characterized by a consistent 0.2–0.4 m thick tillaged clay layer on top of at least 17 m of homogeneous loamy sand, as evidenced by numerous corings at the site and its direct vicinity. This loamy sand overlies a thick aquifer of fluvial sands [*Weerts et al*., 2005]. The regional groundwater gradient drives the exfiltration of brackish to saline groundwater (around 5 g/L Cl), salinized during marine transgressions around 5000 year B.C. [*Post et al*., 2003; *Delsman et al*., 2014b]. The annual precipitation surplus ensures the development of a shallow rain water lens [*De Louw et al*., 2011] on top of the upward flowing brackish groundwater flow, enabling the cultivation of freshwater-dependent crops.

In our monitoring setup, we physically separated tile drain and ditch discharge and recorded their flow rate and electrical conductivity, referenced at 25°C (EC25), at 15 min intervals during two measurement periods (30 May 2012 to 1 December 2012 and 15 April 2013 to 1 October 2013). A combined water, salinity, and heat balance approach was used to separate the groundwater component (exfiltration and infiltration) from ditch discharge measurements; uncertainty was assessed using Monte Carlo analysis [*Delsman et al*., 2014a]. Meteorological information was recorded at a station in the south-western end of the agricultural field, and groundwater heads and EC25s were measured in several dual piezometers (screened at 0.8–1.0 and 1.8–2.0 m BGS) in a transect perpendicular to the ditch, both on and between tile drains. A piezometer in the center of the ditch was screened at 2.8–3.0 m depth. Soil moisture sensors were placed at different depths both on and between tile drains. We installed eight temperature sensor arrays in transects both perpendicular and parallel to the ditch-field interface, each equipped with 10 thermistors spaced at 35 cm intervals to monitor groundwater temperature variations caused by seasonal and diurnal surface temperature variations. The groundwater salinity distribution was evaluated using geophysical surveys before and after the measurement period. Potatoes and lettuce were grown on the field for the first and second year of study, respectively. Conditions were generally wet during the 2012 growing season, 2013 was characterized by a prolonged dry period from June to August. While the measurement setup should have ensured constant water levels in the ditch, a pump malfunction caused the ditch to run dry on 8 July 2013, conditions were restored on 31 July [*Delsman et al*., 2014a].

2.2. Modeling Approach

We used SEAWAT [*Langevin et al*., 2008] to set up a variable-density groundwater flow and transport model of a subsection of the studied field site. The model extends from a tile drain to the midpoint between two adjacent tile drains, and from the midpoint of the agricultural field to the midpoint of the road on the other side of the investigated ditch (Figure 1). Model extent and cell dimensions are listed in Figure 1. The perpendicular positions of the ditch and tile drain ensure a decidedly 3-D flow pattern and prohibit the use of a vertical-2-D schematization. Lateral boundaries were chosen to represent shallow groundwater divides and modeled as no-flow boundaries. The chosen model domain and lateral boundaries were deemed allowable even at depth, as this pattern of tile drains and perpendicular ditches is repeated successively over an area of about 10 km^2 around the field site, resulting in a predominantly vertically upward regional groundwater

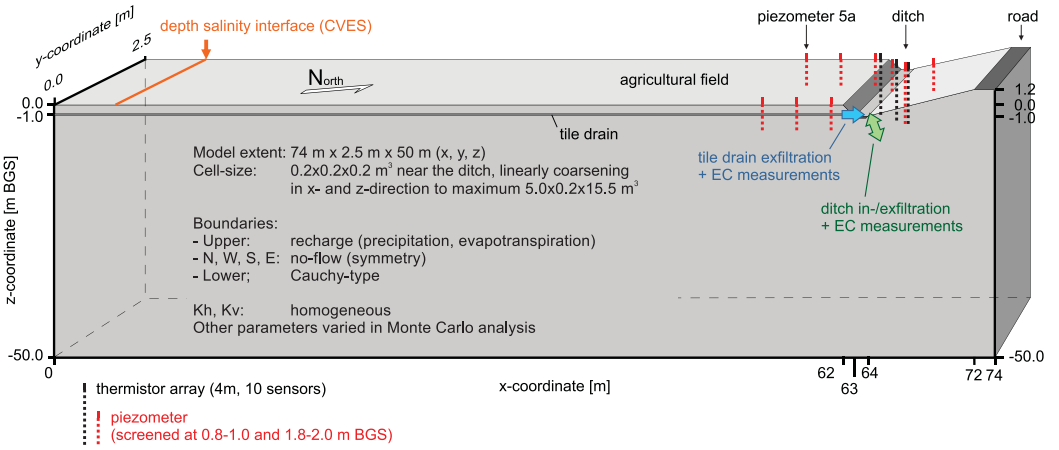

Figure 1. Conceptual model representation and approximate measurement locations.

flow component. Similar model conceptualizations have been widely used to study shallow groundwater flow in similar Dutch polder settings [e.g., *Devos et al*., 2002; *De Louw et al*., 2011, 2013; *Eeman et al*., 2011, 2012]. The lower boundary was chosen deep enough to not influence flow paths to ditch and tile drains, and was modeled as a Cauchy boundary condition, using heads measured in the underlying aquifer at a representative piezometer 500 m to the northwest. The ditch was modeled using a RIV boundary condition, allowing the infiltration of extraneous diverted freshwater, while the tile drain was modeled using a DRN boundary condition. Even though subsurface heterogeneity is widely regarded as a key uncertainty in groundwater modeling [*Refsgaard et al*., 2012], we assumed a homogeneous subsurface in our model. We considered this assumption allowable given the relatively uniform loamy sands that are present to a significant depth, as evidenced from numerous corings in the area [*Van der Meulen et al*., 2013], the relatively uniform apparent resistivity pattern measured using electrical resistance tomography [*Delsman et al*., 2014a], and the shallow flow system usually associated with narrowly spaced tile drains [*Hooghoudt*, 1940]. We further assumed subsurface heterogeneity below the uniform loamy sands to be, given the upward flow at this depth, implicitly accounted for in the lower boundary conductance during the conditioning process (see section 2.3). The thin superficial clay layer was not included in the model as measured groundwater levels were always situated below this clay layer.

The simulated period ranged from 1 May 2012 to 1 October 2013, concurrent with available measurements [*Delsman et al*., 2014a]. The subsurface solute distribution was initialized using a steady state starting condition, followed by a spin-up period of 5 years preceding the model period. The spin-up period was chosen in accordance with previous modeling in similar settings [*De Louw et al*., 2013] and was checked by inspection of the established model saltwater interface. Model stress periods were 1 day, with adaptive time stepping applied in transport modeling. Forcing data for the spin-up period were obtained from nearby meteorological stations operated by the Royal Netherlands Meteorological Institute, and forcing data for the analyzed period were measured by the local meteorological station [*Delsman et al*., 2014a]. Evapotranspiration (ET) was calculated using the FAO Penman-Monteith dual crop-coefficient method, with growing stages based on weekly visual observations, and potential evapotranspiration corrected to actual using shallow soil moisture data [*Delsman et al*., 2014a].

Using the SEAWAT model code precluded the inclusion of the unsaturated zone in our model structure. We still preferred SEAWAT over arguably more physically realistic model codes as, e.g., HydroGeoSphere [*Therrien et al*., 2005], because of much lower calculation times. Moreover, we assumed the influence of the unsaturated zone to be minor, given that the shallow local groundwater levels and widespread occurrence of macropores would result in little delay of recharge to the water table. Temperature-corrected electrical conductivity of groundwater was assumed to behave conservatively and mix linearly, acceptable for the field conditions considered [*Delsman et al*., 2014a]; EC25 was therefore modeled as a conservative species, linearly influencing groundwater density [*Post*, 2012]. EC25 of precipitation was set to 0.2 mS/cm, EC25 of regional groundwater flow to 21.8 mS/cm, based on available measurements. EC25 of surface water during

infiltration was set to measured ditch EC25, averaging around 1.0 mS/cm during the August 2013 infiltration period [*Delsman et al*., 2014a]. We used a dual-domain approach to simulate the dispersion of solutes, as dispersion appears to be mainly related to mass transfer between a relatively mobile phase of the porous medium and stagnant pores and small low-permeable regions [*Lu et al*., 2009; *De Louw et al*., 2013; *Molz*, 2015]. With the dual-domain approach accounting for the main dispersion mechanism, we set the longitudinal dispersion coefficient to a suitably low value of 0.01 m, and assigned a molecular diffusion coefficient of 10^{-9} m^2 s^{-1} to all model cells. We invoked the rewetting capability of SEAWAT to allow the groundwater level to fluctuate across model layers, and counteracted occurring stability problems by decreasing respective model time steps in case of nonconvergence.

Although interest at the Schermer field site was focused on the transport of salts toward surface water, we additionally modeled groundwater temperature to enable model conditioning using observations of groundwater temperature data. We modeled temperature using a decoupled approach, instead of including temperature as a separate species in the SEAWAT calculation. The decoupled approach, applying MT3DMS [*Zheng*, 2009] with the SEAWAT-derived groundwater flow solution, enabled us to specify a fixed temperature boundary to the uppermost model cells, even though the vertical location of the uppermost saturated cells varied due to rewetting of model cells. We considered the effect of occurring temperature variations (2–16°C) on groundwater density to be negligible, and daily stress periods were considered small enough to neglect unsaved flow variations due to density variations within stress periods [*Langevin et al*., 2008]. For temperature modeling, we did not apply a dual-domain approach, but used the analogy between heat and solute transport as outlined by *Langevin et al*. [2008]. Temperature values for the upper boundary were derived from measured temperature values around groundwater levels for the measurement period, and a rolling weekly average of air temperature obtained from nearby meteorological stations for the spin-up period. The lower boundary condition was set to a fixed temperature of 10.63°C, the long-term average temperature in the area. Minor alterations to SEAWAT and MT3D proved necessary to handle specifics of our modeling approach, these are outlined in supplementary information.

2.3. Uncertainty Evaluation

We evaluated uncertainty in our modeling approach by applying the GLUE methodology of *Beven and Binley* [1992]. GLUE recognizes that, given uncertainties and (often epistemic rather than random) errors in model structure, model parameterization, and observational data, multiple models will be equally good descriptors of reality and thus exhibit equifinality [*Beven*, 2006]. Rather than trying to optimize a single parameter set for a given model structure, GLUE retains multiple model structures or model parameterizations that are considered behavioral given some (subjective) adequate fit to available measurement data. Results of all behavioral models are then weighted according to a likelihood measure (be it formal, informal or fuzzy), expressing a degree of confidence in the model (or parameter set). The prior collection of models is generally obtained by simple Monte Carlo sampling of parameter ranges, although more elaborate Markov Chain Monte Carlo methods have also been used [e.g., *Blasone et al*., 2008; *Rojas et al*., 2010]. A more complete description of GLUE is presented by *Beven and Binley* [1992] and *Beven* [2006a, 2009]. GLUE accounts for errors in model structure, parameters, or observations implicitly, as these errors are characterized by the spread in the collection of behavioral models. GLUE therefore does not require an a priori defined error structure [*Beven*, 2009]. The GLUE methodology has been criticized for lacking the rigor and objectivity of formal statistical approaches, particularly for the subjectivity involved in the distinction between behavioral and nonbehavioral models [*Mantovan and Todini*, 2006; *Stedinger et al*., 2008; *Clark et al*., 2011]. Despite this controversy, the GLUE methodology has found widespread use in environmental modeling [*Beven and Binley*, 2013].

In our approach, we used a Latin Hypercube sampler (LHS) to uniformly sample parameter ranges of 13 model parameters, either in linear or in log-space (parameters and sampled ranges in Table 2). Parameter ranges were derived either from field measurements, from exploratory modeling or based on literature ranges. Model run times (around 1 h per run) necessarily limited the total number of Monte Carlo runs to 10,000, a small sample given the 13-dimensional parameter space (already requiring 10^{13} runs to evaluate all parameter combinations when each parameter range is subdivided in ten parts). LHS is, however, more efficient in representatively sampling the entire parameter space than ordinary random sampling, and has been shown to require only about 10% of samples compared to ordinary sampling to obtain representative uncertainty estimates [*Gwo et al*., 1996; *Yu et al*., 2001]. Although perhaps contrary to the common practice of separating calibration and validation data sets [e.g., *Foglia et al*., 2009], we used the entire available

measurement period both for conditioning parameter estimates and evaluating the conditioned result, in order to utilize the maximum of available information. Also again note that we did not explicitly include model structural error associated with conceptualizations of the subsurface in our analysis [e.g., *Foglia et al*., 2013], even though the uncertainty in conceptualizing the subsurface is widely regarded as the key source of uncertainty in groundwater models [*Refsgaard et al*., 2012]. This paper aims only to demonstrate the influence of disinformation in observational data in model conditioning using different data types, it makes no attempt to disentangle the different causes (errors in model structure of course being a prominent one) of this disinformation. In addition, we assumed geologic uncertainty to be low in our small, extensively monitored model domain for reasons explained previously.

2.4. Conditioning of Parameter Sets to Different Targets

2.4.1. Conditioning Using Different Observational Data Types as Targets

We compared the value of different observational data types by conditioning model parameters using a number of informal likelihood measures. Each of these was based on a different type of observational data. In the synthetic variant of the model (see section 2.4.2), observations were constructed from model results with added measurement noise; in the real-world analysis (see section 2.4.3), the actual observations were used. As transformations of observations have been shown to differently condition model parameters [*Carrera et al*., 2005; *Vandenbohede and Lebbe*, 2010; *Rasa et al*., 2013], we included both "raw" observations and integrative transformations thereof (exfiltration and salinity load cumulatives and temperature envelopes). The types of observational data or conditioning targets considered were (see also Figure 1, italic abbreviations are used in Figures 3 and 6, n denotes number of observations):

1. Heads at 14 piezometers in the field and the ditch (*H*, n = 14 × 356),
2. Total exfiltration to tile drains and exfiltration and infiltration to/from the ditch (*Q di* + *dr*, n = 2 × 356),
3. Cumulative exfiltration to tile drains and ditch (*Q cum*, n = 2 × 356),
4. EC25 of tile drain exfiltration and EC25 of ditch exfiltration (*EC di* + *dr*, n = 2 × 356),
5. Cumulative salinity load of tile drains and ditch (*S cum*, n = 2 × 356),
6. Fitted depth of salinity interface on 29 March 2012 (first geophysical survey) (*D iface*, n = 1),
7. Temperature variation at eight different depths at three locations around the ditch (*T var*, n = 3 × 8 × 356),
8. Fitted envelope of temperature amplitude and phase at three locations around the ditch (*T env*, n = 3 × 2).

All likelihood measures were of the form:

$$\boldsymbol{L}(O_y|\theta_k) = C\left\{\frac{1}{m}\sum_{j=1}^{m}\left[W_j\sum_{i=1}^{n}(\hat{y}_{i,j,k}-y_{i,j})^2\right]\right\}^{-1}; \quad (1)$$

evaluating the likelihood L of simulating the observations O of type y given the kth model (parameterization) θ. The likelihood is based on the weighted average over m contributing observational series of the means of squared residuals $(\hat{y}_{i,j,k}-y_{i,j})$, with $y_{i,j}$ the ith observation of series j of observation type y and $\hat{y}_{i,j,k}$ the respective prediction by model k. W_j is a weight based on the reciprocal of the average sum of squared errors for series j (see Table 1), C is a scaling constant to ensure the cumulative likelihood is one. For each measure, we considered the bottom 95% as nonbehavioral and discarded these runs from further analysis. We used this relative, rather than absolute, criterion to enable intercomparison between equal numbers of behavioral runs. Drawback of this relative approach is the loss of a clear relation between measurement error (clearly different between different observational data types) and the cutoff criterion. For heads (*H*), exfiltration (*Q di* + *dr*), EC25 (*EC di* + *dr*), and temperature variation (*T var*), the residuals between measurements and model results were calculated on the time series, and cumulatives (*Q cum* and *S cum*) were calculated after first discarding missing periods in the measurement series. The depth of the salinity interface on 29 March 2012 (first geophysical survey) (*D iface*) was evaluated after fitting a cumulative normal distribution to both the measured and modeled salinity distribution. A cumulative normal distribution has been shown to adequately characterize the mixing zone between fresh and saline water [*De Louw et al*., 2011]. For the modeled and observed salinity interfaces, $(\hat{y}_{i,j,k}-y_{i,j})^2$ in equation (1) is the squared difference between the first moments of both fitted cumulative normal functions, signifying the centers of the mixing zones between fresh and salt water. We also fitted a sine function to both measured and modeled groundwater temperature [*Stallman*, 1965; *Vandenbohede et al*., 2014]:

Table 1. Overview of Metrics Used in This Work and Their Purpose

Metric	Purpose
$\boldsymbol{L}(O_y\|\theta_k)=C\left\{\frac{1}{m}\sum_{j=1}^{m}\left[W_j\sum_{i=1}^{n}(\hat{y}_{i,j,k}-y_{i,j})^2\right]\right\}^{-1}$	Likelihood function in GLUE analyses. The likelihood function is the reciprocal of the weighted average over the sum of squared errors of m observational data series j
$W_j=\overline{SSE}_j^{-1}=\left(\frac{1}{o}\sum_{k=1}^{o}\sum_{i=1}^{n}(\hat{y}_{i,j,k}-y_{i,j})^2\right)^{-1}$	Weight in GLUE likelihood function, the reciprocal of the sum of squared errors of observation data series j averaged over all o model runs k
$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}$	Root-mean-squared error, used as a metric to evaluate conditioned model results to a priori model results, and to observations
$\mathrm{NRMSE}=\frac{\mathrm{RMSE}}{(y_{max}-y_{min})}\cdot 100\%$	Normalized root-mean-squared error, used as a metric to evaluate conditioned model results to observations. The NRMSE presents the RMSE (model misfit) as a percentage of the range in observations

$$T(t)=T_m+A\sin\left(\frac{2\pi}{365}(\varphi+t)\right), \quad (2)$$

in which $T(t)$ is temperature (°C) at time t (day), T_m is mean temperature (°C), A is amplitude (°C), and φ is phase (day) . We then used residuals for the fitted values of both amplitude and phase at thermistor locations in the likelihood calculation (*T env*).

2.4.2. Synthetic Model Experiment

We studied the potential value of different data types in conditioning a groundwater flow and transport model by conducting a synthetic model experiment, based on the real field site. We state potential value, as the synthetic experiment excludes possible confounding effects of model structural errors and non-Gaussian observational errors (Gaussian measurement error is, however, included). In the synthetic model experiment, observations were generated from a forward model run. We used the approximate centers of the various Monte Carlo parameter ranges as parameter values in the forward model run (Table 2). The generated observations were perturbed with random, zero-mean Gaussian noise to include observational uncertainty. Noise variance was based on reported measurement uncertainty of the different sensors employed in the field, values are listed in Table 3.

We studied the marginal posterior parameter distributions to observe conditioning patterns of the different model parameters to the different conditioning targets. We further compared the improvement of the fit for each differently conditioned median model outcome (root-mean-squared error (RMSE), Table 1) to the prior estimate, i.e., the median of the entire prior set of models, and compared the width reduction of the 5–95% uncertainty interval for each differently conditioned model outcome compared to the prior estimate. We plotted these results in radar plots, with on the different axes the different outcome measures, and colored lines each representing model results conditioned on a different target.

We finally employed hierarchical cluster analysis (HCA) to reveal patterns in model conditioning to the different conditioning targets. HCA groups samples into significantly different clusters, according to the,

Table 2. Selected Parameters in Monte Carlo Analysis and Sampled Ranges, and Parameter Values in Forward Run Used for Synthetic Case

Name	Symbol	Min	Max	Fwd Run	Unit	Distribution
Horizontal hydraulic conductivity	K_h	0.1	10	1.0	m d^{-1}	Log-uniform
Specific yield	Sy	0.01	0.2	0.1		Uniform
Vertical anisotropy ratio	K_h/K_v	1	20	10		Uniform
Hydraulic resistance tile drain	C_{drain}	0.1	2	0.5	day	Log-uniform
Hydraulic resistance ditch	C_{ditch}	0.1	2	0.5	day	Log-uniform
Hydraulic resistance lower boundary condition	C_{lbc}	100	1×10^5	2.0×10^3	day	Log-uniform
ET adjustment factor potato	ETF_P	0.5	1.5	1.0		Uniform
ET adjustment factor lettuce	ETF_L	0.5	1.5	1.0		Uniform
ET adjustment factor bare soil	ETF_B	0.5	1.5	1.0		Uniform
Total porosity	θ_{tot}	0.4	0.7	0.55		Uniform
Immobile fraction porosity	$f_{\theta im}$	0.2	0.7	0.4		Uniform
Mass transfer coefficient	ζ	1×10^{-4}	1	1×10^{-3}	day^{-1}	Log-uniform
Thermal conductivity of the porous medium	k_t	1.5	3.5	2.5	W m^{-1} K^{-1}	Uniform

Table 3. Measurement Uncertainty Applied as Noise to Synthetic Observations

Name	Device	Measurement Uncertainty[a]	Reference
Heads	Schlumberger Diver	±0.002 m	*Schlumberger* [2010]
Exfiltration to tile drains	Itron propeller flow meter	±1.5%	*Barfuss et al.* [2011]
In/exfiltration to and from ditch	Itron propeller flow meter	±1.5%	*Barfuss et al.* [2011]
EC25 of tile drain and ditch exfiltration	Schlumberger CTD-Diver	±1.0%	*Schlumberger* [2010]
Concentration profile 29 March 2012	ABEM Terrameter	±5.0%	*Dahlin and Loke* [1998]
Temperature	Onset S-THB	±0.2°C	*Onset* [2013]

[a]Uncertainty is reported as 2 × (relative) standard deviation.

usually Euclidean, distance between groups [*Anderberg*, 1973]. HCA is, for instance, used extensively in geochemistry, where clustering is often used to identify distinct geochemical groundwater types from water samples [e.g., *Thyne et al*., 2004; *Guggenmos et al*., 2011]. In our analysis, HCA samples consisted of either one (behavioral) or zero (not behavioral) for each run for each conditioning target. Sets of behavioral runs conditioned to different targets that share many behavioral runs will be quickly grouped together (smallest Euclidean distance), while sets that overlap the least will be furthest apart.

2.4.3. Conditioning to Real Field Observations

After the synthetic model experiment, we repeated the conditioning procedure but now using the real field observations. Results were analyzed as for the synthetic experiment. Using real observations, the conditioning process will be subject to possibly confounding effects of epistemic errors in both model structure and observational data. We compared conditioning results for the real and synthetic observations to yield insight in the effects of epistemic errors in both model structure and observational data on the conditioning process. Additionally, we visually evaluated differently conditioned model fits to observations to identify time-variant patterns in model fits and conditioning behavior that might be indicative of model error [*Hill and Tiedeman*, 2007]. Note that the sequential evaluation of parameter sensitivity to model-generated and real-world measurements somewhat resembles the sequential use of fit-independent and fit-dependent statistics [*Hill and Tiedeman*, 2007; *Foglia et al*., 2009], and the a priori and a posteriori identifiability of parameters investigated by *Dobre et al*. [2012] .

2.5. Model Conditioning to a Combination of Different Observations

In a final step, we conditioned our model using a combination of conditioning targets, to investigate whether a jointly conditioned model can satisfactorily reproduce groundwater flow and solute transport at the Schermer field site. From the analysis using the real observations, we selected the three best-performing data types. Joint conditioning was achieved by successive Bayesian updating of likelihoods (multiplying likelihoods and rescaling the sum to one).

We compared the results of the jointly conditioned model to model results when conditioned on head data alone (again from the real observations analyses). The problems of parameter nonuniqueness when calibrating to head data alone have been repeatedly emphasized in literature [*Weiss and Smith*, 1998; *Barth and Hill*, 2005a,b; *Hunt et al*., 2006; *Hill and Tiedeman*, 2007]. However, calibrating to only head data is still common practice in groundwater modeling, owing to the fact that heads are the most readily, and often the only, available observational data. We evaluated the resulting fits of model results to observations using the RMSE and RMSE normalized to the range of observations (NRMSE) (Table 1).

3. Results

Despite an additionally implemented adaptive time stepping routine, 1938 of the 10,000 SEAWAT model runs failed due to nonconvergence, a common issue with the rewetting option in MODFLOW models [e.g., *Doherty*, 2001]. Note that MODFLOW (but not SEAWAT) now offers a Newton-based solver to avoid rewetting nonlinearities [*Niswonger et al*., 2005], and a defined saturated thickness in the top layer has been shown to work well for many cases where the change in saturated thickness relative to the total flow depth is small [*Sheets et al*., 2014]. In the subsequent MT3DMS temperature calculation, a further 166 model runs ended prematurely, leaving 7896 model runs for further evaluation.

Table 4. Cutoffs Associated With Relative Top 5% Limit-Of-Acceptability for Synthetic and Real-World Cases, Reported as RMSE Values

Conditioning Target	Synthetic Experiment	Real Observations	Unit
Heads at 14 piezometers in the field and the ditch	0.032	0.061[a]	m
Exfiltration to tile drains and in/exfiltration to ditch	2.08	3.11	m^3/d
Cumulative exfiltration to tile drains and ditch	144.3[b]	79.6[b]	m^3
EC25 of tile drain and ditch exfiltration	1.95	2.28	mS/cm
Cumulative salinity load of tile drains and ditch	2526.3[b]	800.8[b]	$mS/cm \cdot m^3$[c]
Depth of salinity interface on 29 March 2012	0.14	0.14	*M*
Temperature variation at eight depths, three locations	0.22	0.46	°C
Temp. envelope of amplitude and phase, three locations	(0.09)[d]	(0.30)[d]	

[a]Note that this RMSE is below reported RMSEs in section 3.2, being the average over 14 piezometers.
[b]For cumulative exfiltration and salinity load, the likelihood distribution was much steeper in the synthetic experiment than for the real observations, causing a higher cutoff value.
[c]This unusual unit stems from the fact that we model EC25 as a conservative solute.
[d]This target combines two different units in its calculation (°C and day), reporting an RMSE value therefore makes no physical sense. Note that in the calculation of the likelihood, we normalized the different contributions and thereby allowed different types of data to be combined. We still opted to report RMSEs in this table for increased intelligibility of the other conditioning target values.

3.1. Conditioning Using Different Observational Data Types

3.1.1. Synthetic Model Experiment

The top 5% relative limit of acceptability resulted in 392 behavioral runs for each conditioning target; cutoff values for the different targets are presented in Table 4. We evaluated the conditioning of parameters by comparing the marginal cumulative parameter distributions for each conditioning target (Figure 2). Note that the prior cumulative likelihood diverges for several parameters from a uniform distribution due to non-convergence of tested models in distinct regions of the parameter space (dotted black line versus straight grey line in Figure 2). In general, similar parameters were sensitive to conditioning, regardless of the conditioning target. The most sensitive model parameters were the hydraulic resistance of the lower boundary condition (C_{lbc}) and the hydraulic conductivity (K_h). Less sensitive was the tile drain resistance (C_{drain}), specific yield (Sy), total porosity (θ_{tot}), and the evapotranspiration factors for bare soil and lettuce (ETF_B and ETF_P). The remaining parameters showed little conditioning by the different conditioning targets. Conditioning decreased 5–95 percentile ranges of parameter estimates of the best-conditioned parameters (C_{lbc}, K_h) by 56% and 27% respectively (averaged over the different conditioning types), average reduction for all parameters was only 10%.

The hydraulic resistance of the lower boundary condition (C_{lbc}) was more tightly constrained by the salinity related than the other data types; an expected result, given that the lower boundary is the main source of salt in the model. The transport-related parameters (θ_{tot}, $f_{\theta,\,im}$, and ζ) were not only understandably insensitive to head-exfiltration type data, but appeared also generally insensitive to salinity-related data types. Only temperature data constrained the total porosity (θ_{tot}). Different conditioning targets resulted in some cases to parameters being constrained to different regions of the parameter space. This was most apparent for K_h, C_{drain}, and the evapotranspiration factors. This also meant that not all parameters were constrained to values in accordance with forward run parameters. For K_h, for instance, conditioning to exfiltration of ditch and drain constrained K_h to somewhat lower values than in the forward run (around 0.5 instead of 1.0 m d^{-1}).

For each behavioral set conditioned to an observational data type, we compared the relative RMSE improvement for the different median model outcomes against the prior (i.e., the entire set of model runs) and plotted the relative improvements in a radar plot (Figure 3a). Each axis shows the relative improvement of an outcome measure (RMSE of median of behavioral model results versus observations for each model outcome, relative to RMSE of all runs), lines represent the different conditioning targets. We also compared the average width of the 5–95% uncertainty interval against the prior (Figure 3b). Conditioning to all observational data types unsurprisingly improved median predictions for their respective model outcomes. Most outcomes were also improved by other conditioning targets. Only the depth of the salinity interface was not improved by the targets exfiltration of tile drain and ditch, temperature envelope, and heads. The average width of the uncertainty interval was reduced for almost all model outcomes by all conditioning targets (Figure 3b). Only the uncertainty interval of the temperature envelopes stayed roughly the same when conditioned to cumulative exfiltration and salinity load targets.

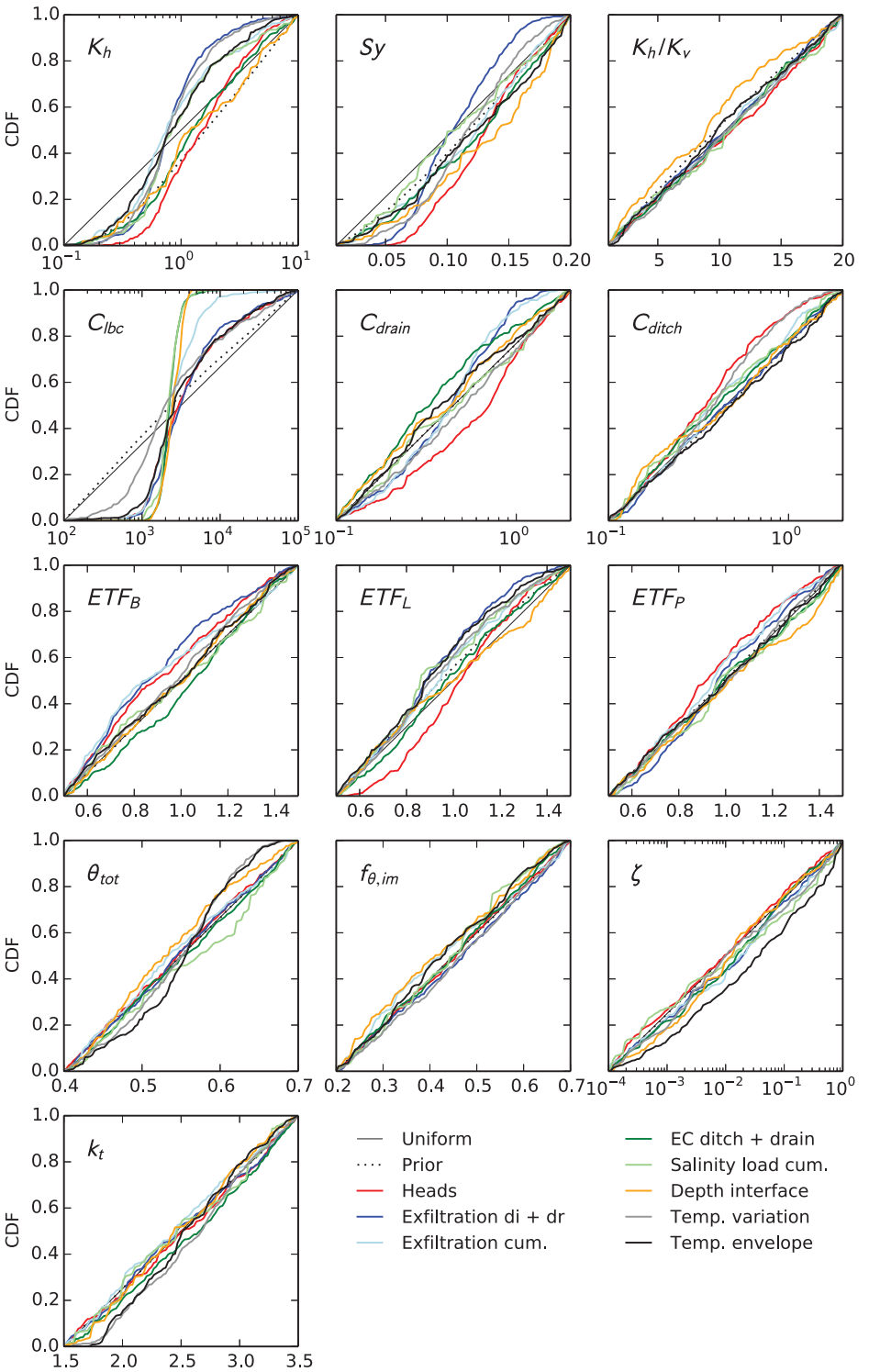

Figure 2. Marginal cumulative density functions for model parameters (Table 2) conditioned on different observational data.

We performed hierarchical cluster analysis (HCA) on the differently conditioned behavioral runs to investigate groupings of different conditioning data. HCA results exposed a logical grouping among the different types of conditioning data (Figure 4). The water quantity data types (exfiltration ditch + drain, cumulative exfiltration and heads) grouped together, in which group exfiltration ditch + drain and cumulative

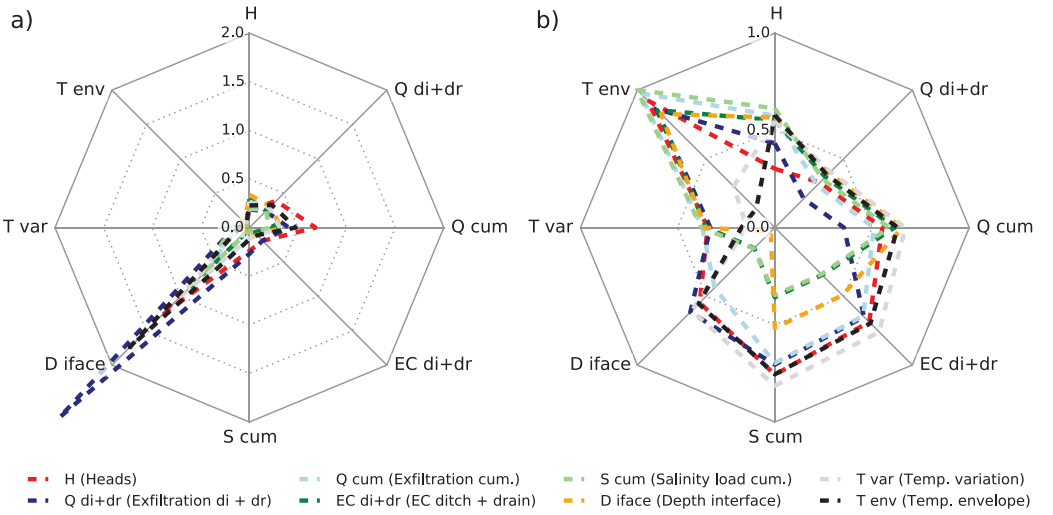

Figure 3. Radar plots of (a) RMSE of conditioned median model predictions relative to prior and (b) average width of conditioned 5–95% uncertainty interval relative to prior for the synthetic experiment. Lines represent model conditioning to diverse measurement data types, radar axes are model outcome measures. Values less than one signify an improvement over the prior, note the difference in scale between (a) and (b).

exfiltration grouped closest together. A similar result was found for the salinity targets, where EC25 and salinity load grouped closest together, and also the temperature data types formed a separate group. We determined the intersection of behavioral runs of the HCA-derived groupings, corresponding to a limits-of-acceptability approach using the different conditioning targets. Remember that the relative rejection criterion of 95% yielded 392 behavioral runs per conditioning target. Results showed that numbers of behavioral runs quickly decreased when different conditioning targets were being combined in subsequent hierarchical groups (Figure 4). While in the main groups still a significant number of parameter sets was behavioral for all group members, only three runs were behavioral for all conditioning targets. Analysis of these three behavioral parameter sets revealed a fairly well-conditioned model (e.g., C_{lbc} values within a factor two, while the initial range varied over 3 orders of magnitude). Note that this analysis of combining conditioning targets is strongly influenced by applying a relative rather than an absolute rejection criterion per target. A

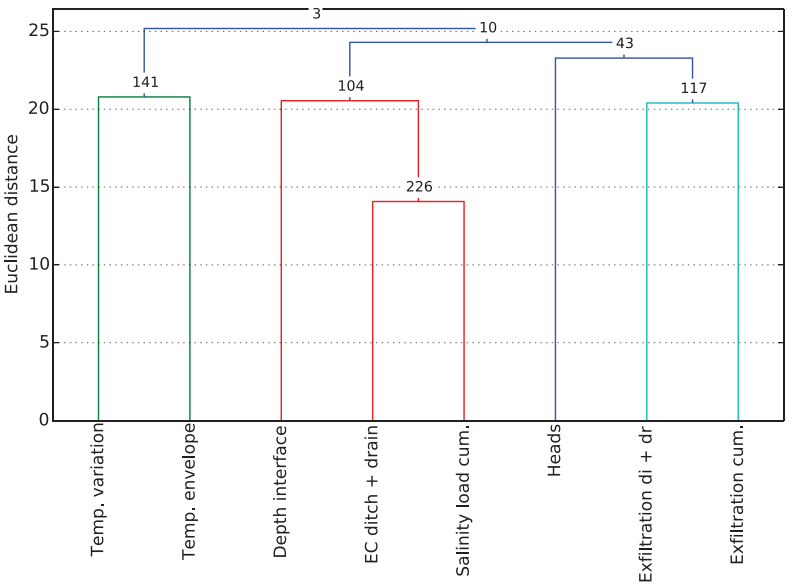

Figure 4. Dendrogram of hierarchical cluster analysis results of behavioral model runs conditioned on different observational data in the synthetic experiment. Numbers at junctions refer to number of behavioral model runs in group intersections.

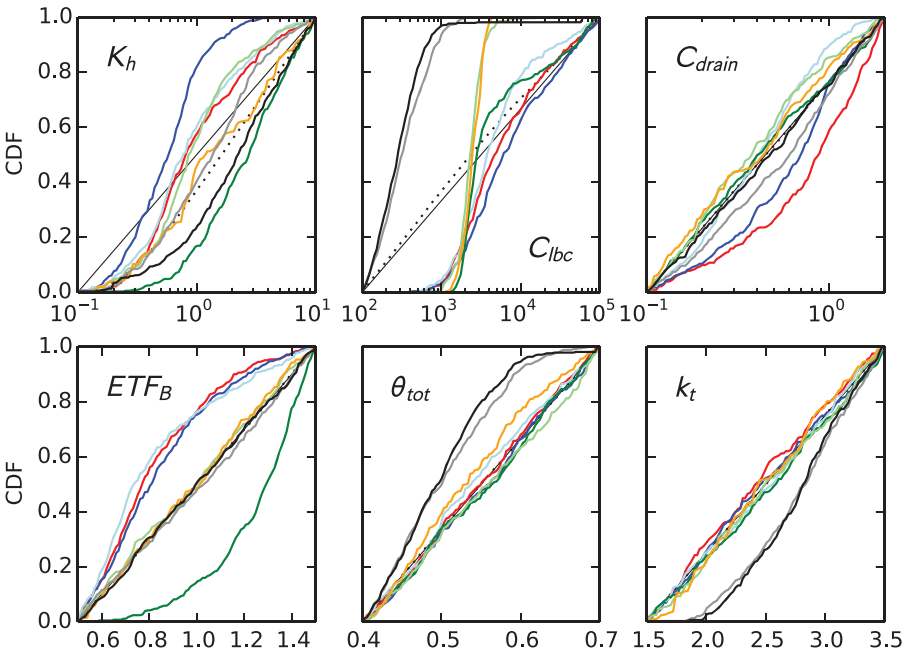

Figure 5. Marginal cumulative density functions for model parameters conditioned on different real observational data. Only parameters showing different conditioning than the synthetic experiment are shown. For legend see Figure 2.

relative rejection criterion does not discriminate between a steeply or more gradually shaped likelihood surface, as an absolute criterion would.

3.1.2. Conditioning to Real Field Observations

We repeated the analyses of the synthetic model experiment but now using the real field observations, instead of model-generated observations (behavioral cutoff values in Table 4). Cutoff values were, due to the larger uncertainty in the real-world case, generally larger than in the synthetic experiment. Similar parameters as in the synthetic experiment were sensitive to conditioning. Marginal parameter distributions that were significantly different from the synthetic experiment (K_h, C_{lbc}, C_{drain}, ETF_B, θ_{tot}, and k_t) are presented in Figure 5. Reduction of 5–95 percentile parameter ranges was similar to the synthetic experiment. Even more apparent than in the synthetic experiment, different conditioning targets often led to parameters being constrained to different regions of the parameter space. Conditioning to temperature data (both direct measurements and temperature envelopes) showed a notably different effect on the constraining of parameters C_{lbc}, θ_{tot}, and k_t. Note that parameters θ_{tot} and k_t are mainly sensitive to temperature data, and differences between conditioning to temperature and other data were expected.

We again compared the relative RMSE improvement for the different median model outcomes against the prior (i.e., the entire set of model runs) for the different real observational data types (Figure 6a). While all conditioning targets improved median predictions for their respective model outcome, changes for other model outcomes were much more varied. Only cumulative salinity load and salinity interface depth improved model predictions for all considered outcomes. Conditioning to the two temperature data types worsened all other model outcomes, while temperature as a model outcome was hardly affected by conditioning to nontemperature targets. While conditioning to the depth of the interface improved all model outcomes, conditioning to other targets significantly worsened the calculation of the interface depth. Some model outcomes (heads and EC25 ditch + drain) were about as well predicted when conditioning to other targets (excluding temperature) than to heads and EC25 ditch + drain themselves. All conditioning targets narrowed the uncertainty interval around the median model result for most different model outcomes (Figure 6b). Transformed data types (cumulatives of exfiltration and salinity load and temperature envelopes) generally constrained similar model outcomes as their untransformed counterparts. However, the cumulatives outperformed their counterparts both in median predictions and in uncertainty interval widths, albeit sometimes slightly. Contrastingly, temperature envelopes mostly performed worse than temperature variation. Comparing these results to those of the synthetic experiment (Figure 3), RMSE improvements

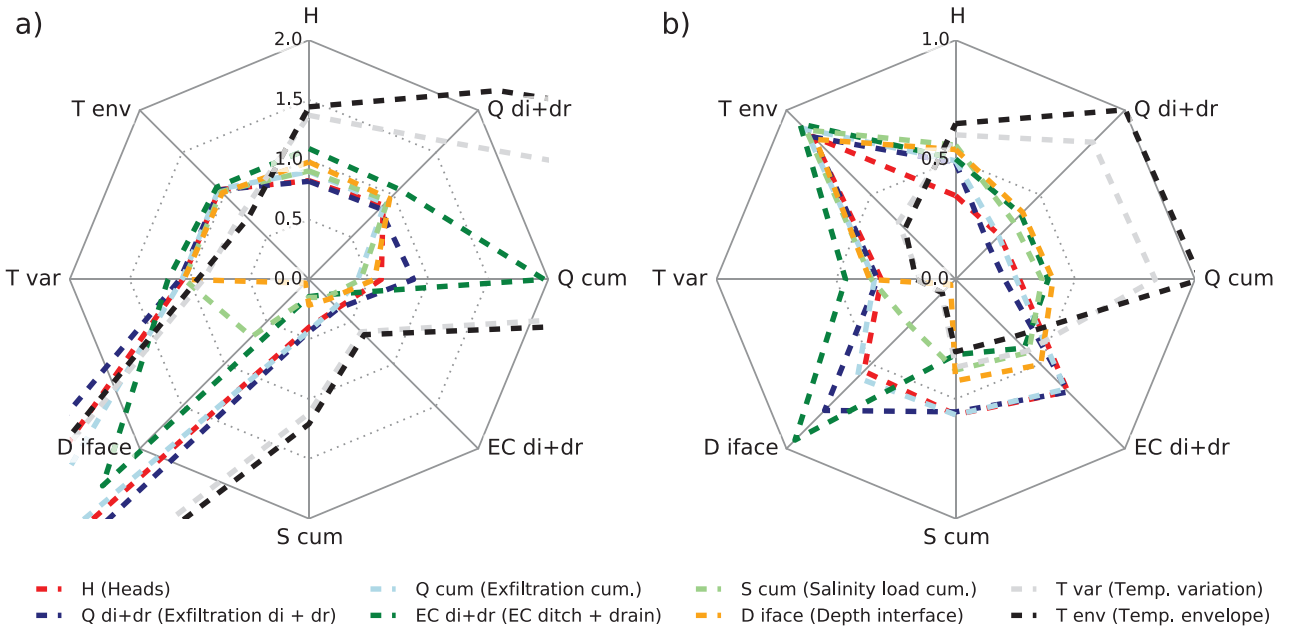

Figure 6. Radar plots of (a) RMSE of conditioned median model predictions relative to prior and (b) average width of conditioned 5–95% uncertainty interval relative to prior for the real observations. Lines represent model conditioning to diverse measurement data types, radar axes are model outcome measures. Note the difference in scale between (a) and (b).

were generally much less, even for the model outcomes, the model was conditioned to. Especially, the temperature targets resulted in far less improvement of RMSEs on other model outcomes. Uncertainty interval widths were generally similar to those of the synthetic experiment.

The worsened agreement in conditioning results between the different targets as observed in the marginal parameter distributions and radar plots was also observed in results of HCA on behavioral runs for the real-world case (not shown). The HCA results showed larger distances and a smaller number of behavioral runs at main group intersections. The grouping differed from the synthetic case and was less intuitive: while the main groups remained the same, cumulatives of exfiltration and salinity load were no longer closest to their respective untransformed counterparts. Where the combination of all targets yielded three behavioral runs in the synthetic experiment, no runs were found behavioral to all conditioning targets in the real-world case.

Visual comparison of differently conditioned median model results to observations (Figure 7) revealed a very close match to observations for almost all of the differently conditioned model runs for both the two head measurements and tile drain and ditch exfiltration (Figures 7a–7d). Only the temperature-conditioned results differed markedly from both the other results and the observations. Model conditioning to heads, flow, and salinity targets clearly all constrained the simulation of the groundwater head gradient and flow field. Temperature measurements, however, did not, and seem to indicate model error in the temperature modeling. In the same visual comparison for the synthetic experiment (not shown), temperature targets did constrain head and flow results to similar values as the other targets. This result therefore points to structural model error in the temperature modeling or is indicative of significant non-Gaussian errors in temperature observations. Matching temperature observations with our current model needed a significant inflow across the lower boundary condition, as evidenced by a low constrained value of C_{lbc}, the consistently higher exfiltration rates and heads, and the increased exfiltration salinities (see below). As our interest at the Schermer field site is not on groundwater temperature, we did not further explore temperature model error.

Differently conditioned model results did, however, differ markedly in their simulation of EC25 of both tile drain and ditch exfiltration (Figures 7e and 7f). Head and flow targets significantly underestimated tile drain and ditch exfiltration EC25. The exfiltration EC25 and cumulative salt load targets better simulated tile drain and ditch exfiltration EC25; depth of the salinity interface performed in-between the heads and flow and

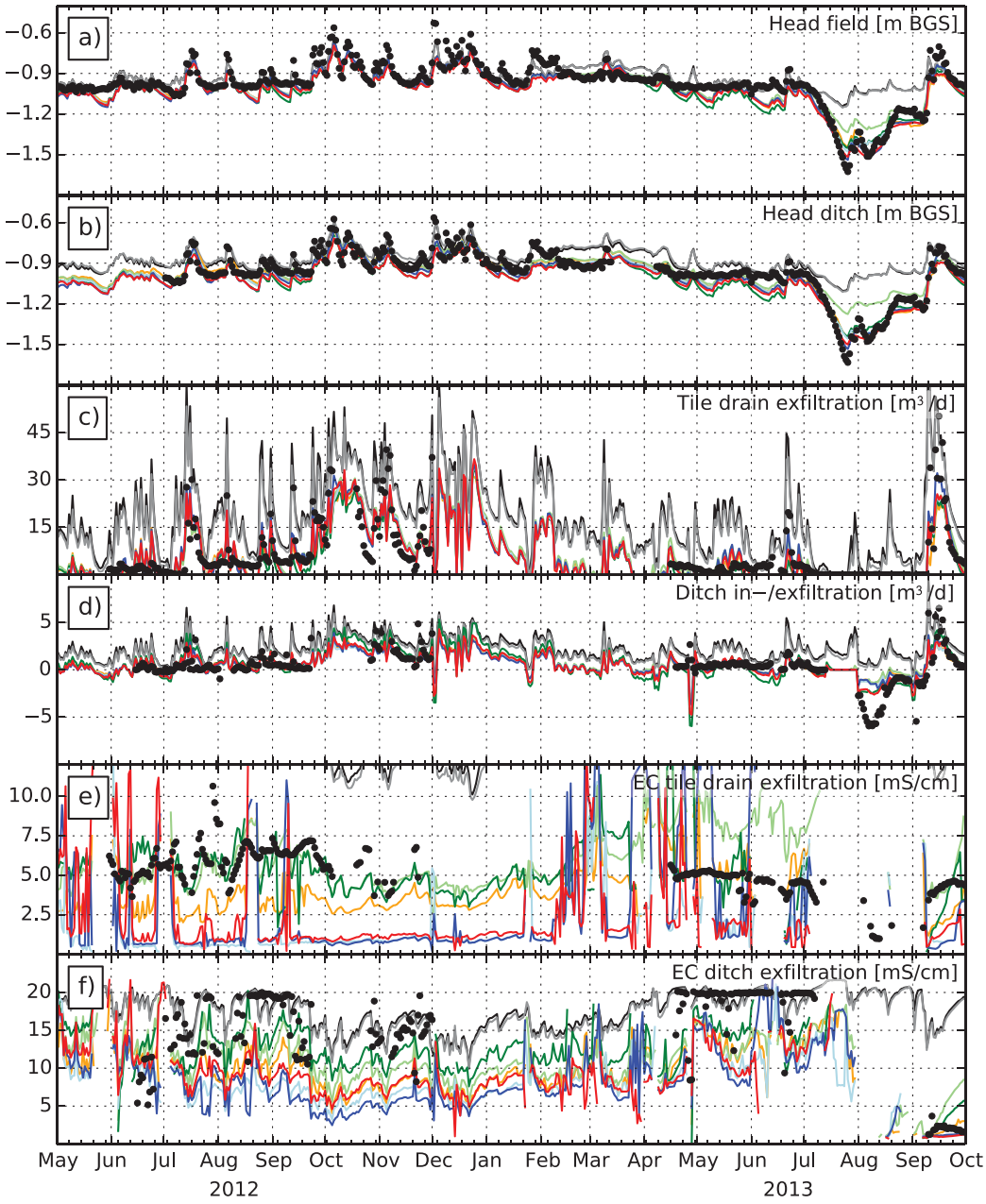

Figure 7. Comparison of differently conditioned median model results to observations for the real-world case, for: (a) head in piezometer 5a (for location see Figure 1), (b) head beneath ditch, (c) tile drain exfiltration, (d) ditch in/exfiltration, (e) EC25 of tile drain exfiltration, and (f) EC25 of ditch exfiltration. Colored solid lines indicate different conditioning targets, colors correspond to colors used in previous figures, black dots denote observations.

the exfiltration salinity targets. Temperature targets significantly overestimated tile drain exfiltration salinity, while ditch exfiltration EC25 was simulated rather well. However, none of the conditioning targets resulted in satisfactory modeling of both tile drain and ditch exfiltration EC25, indicating possible model error. Furthermore, the significant differences between conditioning targets in simulating exfiltration salinity point to the additional complexity of transport modeling. Correct simulation of groundwater flow is required for correct transport simulations, and salinity-related targets therefore constrained the groundwater flow simulation. As evidenced by the diverging differently conditioned model results, the reverse does not necessarily hold, as groundwater transport modeling involves additional processes and parameters that have no influence on groundwater flow.

3.2. Model Conditioning to a Combination of Different Observations

Based on the analysis of conditioning to real observations in section 3.1, we selected three types of conditioning data: (1) drain and ditch exfiltration, (2) cumulative salinity load, and (3) salinity interface depth, as these three types together condition all relevant model outcomes (i.e., all outcomes except temperature). Successive Bayesian updating of model likelihoods (multiplying likelihoods and rescaling the products to sum to one) resulted in 21 behavioral runs out of a possible 7896 (recall that combining all conditioning data would not have yielded any behavioral runs). We compared the obtained jointly conditioned results to results conditioned to only head data (taken from the real-world case of section 3.1). Marginal parameter distributions of jointly conditioned and heads-only-conditioned parameters are given in Figure 8. The marginal parameter distributions of the jointly conditioned model were, due to a lower number of behavioral runs, more discontinuous than the heads-only likelihoods. The distributions showed clear conditioning of parameters K_h and C_{lbc}, parameters K_h/K_v, ETF_P, θ_{tot}, and ζ showed some conditioning. Parameters K_h and C_{lbc} were constrained to different regions of the parameter space (K_h: about 0.4 m/d instead of 0.6 m/d, C_{lbc}: about 2000 days as opposed to around 3000 days) as when conditioned to only head data. While differences for both K_h and C_{lbc} between the heads-only conditioned and the jointly conditioned models are small compared to the wide initial parameter range of 3 orders of magnitude, these differences significantly affect modeled groundwater and solute flow (see below).

Jointly conditioned and heads-only-conditioned model results are presented in Figure 9. The jointly conditioned model head predictions (RMSE 0.09 m, NRMSE 8%) were slightly worse than those of the heads-conditioned model (RMSE 0.08 m, NRMSE 7%) (Figures 9b and 9c); not unexpected as heads were not included in the joint conditioning process. Heads in the jointly conditioned model are a few centimeters above heads in the heads-conditioned results and appear to better match heads during precipitation events. The smaller number of behavioral runs resulted in narrower uncertainty intervals than did the heads-conditioned case, apart from during July and August 2013 when heads fell below the tile drainage level. Measured heads, as the other observations, were not always bracketed by the 5–95% uncertainty interval, a result of the informal likelihood applied in the GLUE methodology [*Beven*, 2006].

Both conditioned models adequately captured the measured fast response of groundwater levels and, consequently, tile drain and ditch exfiltration, to precipitation events. Tile drain exfiltration response for the jointly conditioned and heads-conditioned models was similar, albeit with narrower uncertainty intervals for the jointly conditioned model (Figure 9d). This was reflected in similar RMSE and NRMSE values of tile drain exfiltration (5.4 m^3/d, 11%, versus 5.0 m^3/d, 10%, joint versus heads only). Also ditch in/exfiltration was modeled very similar between the two differently conditioned models (Figure 9e, both 2.5 m^3/d, 6%). However, the prediction of EC25, both of tile drain and of ditch exfiltration, although still far from perfect, clearly improved (Figures 9f and 9g, RMSEs 2.4, 5.8 mS/cm, NRMSEs 27%, 31% for tile drain and ditch exfiltration, respectively, for the jointly conditioned model) and was better constrained than in the heads-conditioned case (RMSEs 4.3, 7.2 mS/cm, NRMSEs 48%, 38%, respectively). Modeled EC25 of tile drainage better resembled observations during the 2012 measurement period than in 2013. Uncertainty in heads and ditch infiltration (Figures 9b, 9c, and 9e) was highest during the period July–August 2013, as heads fell below the tile drain level during this period, and were therefore much less controlled by tile drainage.

4. Discussion

This study evaluated the value of different observational data types in conditioning a groundwater flow, salinity transport, and temperature model of an extensively monitored field site. Adding to previous studies reporting on the value of additional conditioning targets, we applied a computation-intensive global parameter search technique (GLUE). Moreover, we illustrated the possible confounding effects of disinformation in observational data, resulting from errors in model structure and observations, by comparing conditioning in a synthetic experiment to conditioning to real observations.

Our results demonstrated the value of different conditioning data in constraining a field-scale groundwater flow and transport model. The calculation of heads and exfiltration to tile drains and the ditch could be adequately conditioned using all but the temperature observations. However, EC25 of tile drain and ditch exfiltration was only conditioned by salinity-related observations. Including conditioning targets that are sensitive to transport modeling was therefore essential in this case, as has been reported elsewhere [e.g.,

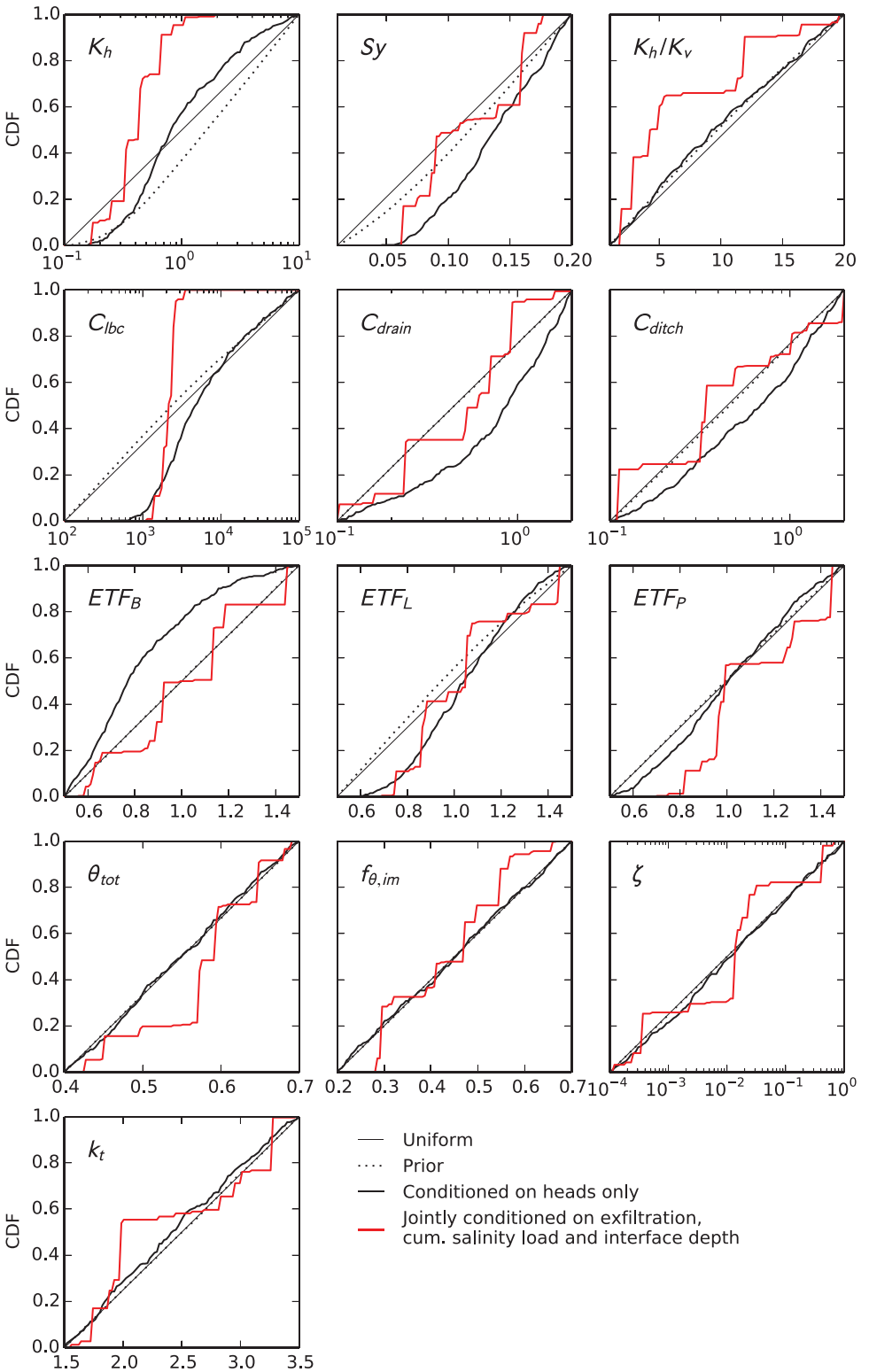

Figure 8. Marginal cumulative density functions for model parameters conditioned on head data (black solid lines), and on a combination of cumulative exfiltration, cumulative salinity load, and salinity interface depth (red solid lines).

Barlebo et al., 1998]. The jointly conditioned model, conditioned jointly on exfiltration, cumulative salinity load, and salinity interface depth, clearly outperformed the heads-only-conditioned model, especially considering salinity-related model targets. At the Schermer field site, a significant part of the flow and almost

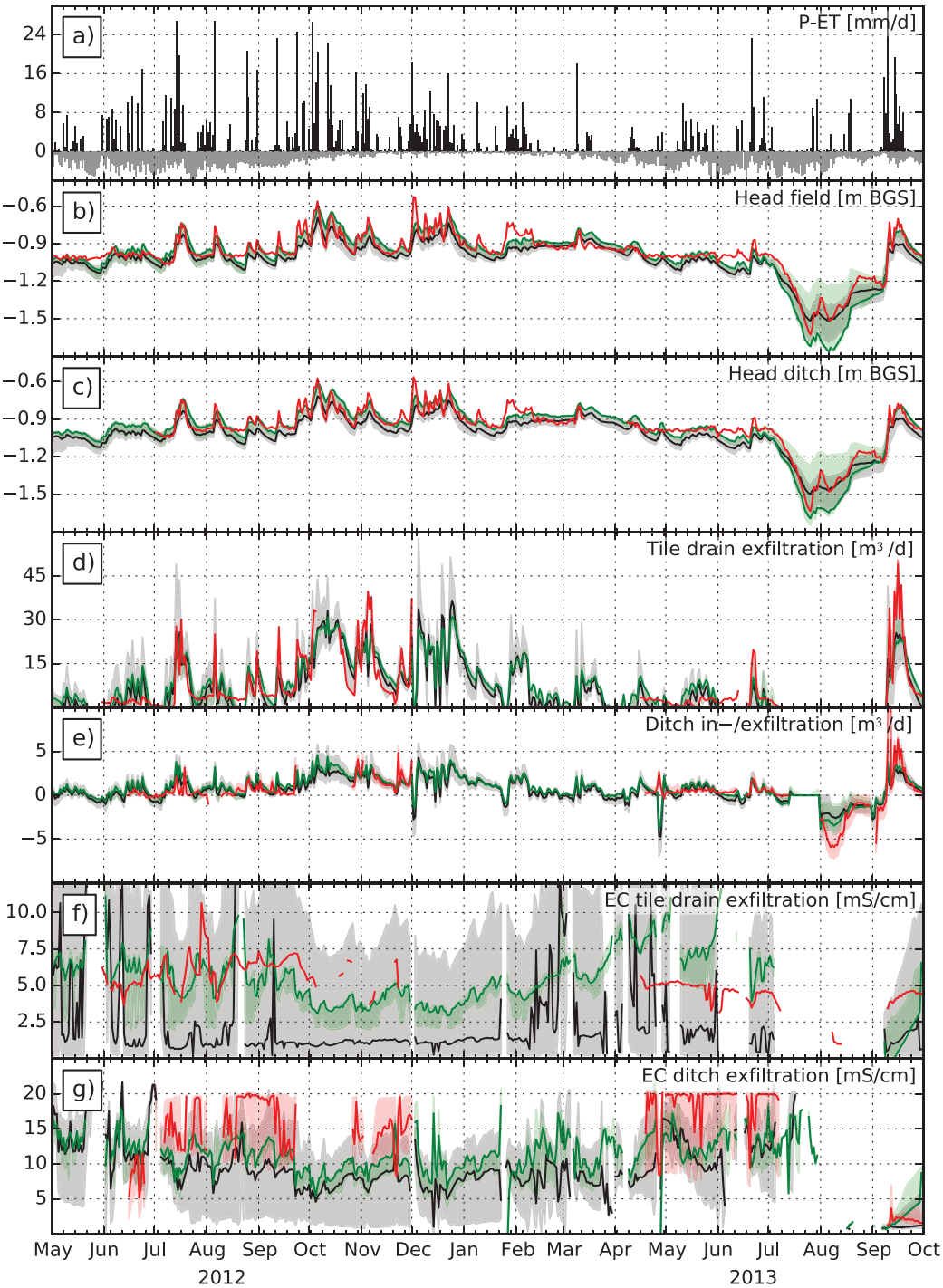

Figure 9. (a) Precipitation and evapotranspiration (mm/d), (b–g) jointly conditioned model result (green, solid line denotes median, shaded area 5–95% prediction interval), heads-only-conditioned model result (black, solid line denotes median, shaded area 5–95% prediction interval), and observations (red) for: (b) head in piezometer 5a (for location see Figure 1), (c) head beneath ditch, (d) tile drain exfiltration, (e) ditch in/exfiltration, (f) EC25 of tile drain exfiltration, and (g) EC25 of ditch exfiltration. Shaded red areas around measured values in Figures 9e and 9g are the 25–75% percentiles of Monte Carlo uncertainty estimates [*Delsman et al.*, 2014a]. Note that the ditch exfiltration peak of 10 September 2013 was likely caused by infiltration-excess overland flow [*Delsman et al.*, 2014a] and was therefore excluded from the conditioning process.

all solutes originate from regional groundwater flow [*Delsman et al.*, 2014a], represented by a Cauchy-type lower boundary condition. Hydraulic resistance of this lower boundary condition was therefore sensitive to all, but especially the salinity-related, conditioning data. The depth of the salinity interface, determined by

cheap and easy-to-operate near-surface geophysics, proved about equally able to improve salinity-related model predictions than laborious, direct measurements of exfiltration EC25. We found cumulative exfiltration and salinity load to lead to somewhat better conditioned model predictions than their untransformed counterparts, in line with results reported by *Carrera et al*. [2005] and *Rasa et al*. [2013].

The use of generalized likelihood uncertainty estimation was computation-intensive, but allowed for insight in the uncertainty surrounding the differently conditioned parameter estimates and model results. The global parameter estimation employed in GLUE ensures that results are not conditional on a (possibly local) optimum and linearization around this optimum. More computationally frugal methods (e.g., fit-independent measures dimensionless scaled sensitivities, composite scaled sensitivities, and fit-dependent measures leverage, Cook's D, DFBETAS statistics [*Hill and Tiedeman*, 2007]) are available to answer similar questions regarding the link between different observations and parameter estimation. These methods are linear approximations, but have been shown to work well for reasonably nonlinear groundwater models [*Foglia et al*., 2007, 2009], and have been suggested to complement global sensitivity methods in preliminary stages of model development [*Foglia et al*., 2009]. How these statistics compare to the results presented here, obtained using GLUE analysis, is an interesting question for future work. The distinctly different methodologies underlying the mentioned frugal statistics (weighted nonlinear regression, seeking an optimal model, observational error should be reflected in weighting of the objective function) and our presented results (different models are equally acceptable, acceptable deviations from observations should be reflected in the rejection criterion) complicate a straightforward comparison.

Errors in model structure and non-Gaussian errors in observational data may compromise the information content of observational data [*Beven et al*., 2008]. Comparing conditioning to real measurement data to the conditioning of a synthetic variant, assuming error-free model structure and only Gaussian errors in observations, indeed pointed to the presence of disinformation in the observations of the real-world case. Parameters were generally more tightly constrained and constraining showed more agreement amongst conditioning data types in the synthetic case. Moreover, where in the synthetic experiment all conditioning targets significantly improved prior estimates for all but one model outcome, improvements against the prior in the real-world case were only minor or even negative. While conditioning to temperature data proved effective in conditioning groundwater flow in the synthetic experiment, this data type was clearly disinformative in the real-world case. Conditioning to temperature data in the real-world case yielded parameter estimates that differed significantly from other conditioning targets. This likely points toward structural errors in temperature modeling, although small observational errors in temperature measurements can already result in important differences between derived and occurring groundwater flow [*Vandenbohede and Lebbe*, 2010]. Overall, these results show that synthetic modeling examples [e.g., *Feyen et al*., 2003; *Hendricks Franssen et al*., 2003; *Rojas et al*., 2010] may overstate the value of different conditioning data, and care must be taken in translating results to real-world models.

In conclusion, while our results showed disinformation in observational data (due to errors in model structure and observations) to confound model conditioning, additional salinity-related observations proved to be essential to adequately condition a field-scale groundwater flow and transport model.

Acknowledgments
Data for this paper may be obtained from the corresponding author. This work was carried out within the Dutch Knowledge for Climate program. Vandenbohede was supported by the Fund for Scientific Research – Flanders (Belgium) where he was a postdoctoral fellow during the time of this research. We thank two anonymous reviewers for their constructive comments that helped to significantly improve this paper.

References

Anderberg, M. R. (1973), *Cluster Analysis for Applications*, Academic Press, N. Y.
Anderman, E., M. Hill, and E. Poeter (1996), Two-dimensional advective transport in ground-water flow parameter estimation, *Ground Water*, *34*(6), 1001–1009.
Barfuss, S., M. Johnson, and M. Neilsen (2011), *Accuracy of In-Service Water Meters at Low and High Flow Rates*, Water Res. Found., Denver, Colo.
Barlebo, H., M. C. Hill, D. Rosbjerg, and K. Jensen (1998), Concentration data and dimensionality in groundwater models: Evaluation using inverse modelling, *Nord. Hydrol.*, *29*(3), 149–178.
Barlebo, H. C., M. C. Hill, and D. Rosbjerg (2004), Investigating the Macrodispersion Experiment (MADE) site in Columbus, Mississippi, using a three-dimensional inverse flow and transport model, *Water Resour. Res.*, *40*, W04211, doi:10.1029/2002WR001935.
Barth, G., and M. C. Hill (2005a), Numerical methods for improving sensitivity analysis and parameter estimation of virus transport simulated using sorptive-reactive processes, *J. Contam. Hydrol.*, *76*(3-4), 251–77, doi:10.1016/j.jconhyd.2004.10.001.
Barth, G., T. Illangasekare, M. C. Hill, and H. Rajaram (2001), A new tracer-density criterion for heterogeneous porous media, *Water Resour. Res.*, *37*(1), 21–31.
Barth, G. R., and M. C. Hill (2005b), Parameter and observation importance in modelling virus transport in saturated porous media-investigations in a homogenous system, *J. Contam. Hydrol.*, *80*(3-4), 107–29, doi:10.1016/j.jconhyd.2005.06.012.

Beaujean, J., F. Nguyen, F. Kemna, A. Antonsson, and P. Engesgaard (2014), Calibration of seawater intrusion models: Inverse parameter estimation using surface electrical resistivity tomography and borehole data, *Water Resour. Res.*, *50*, 6828–6849, doi:10.1002/2013WR014020.

Beven, K. J. (1989), Changing ideas in hydrology—The case of physically-based models, *J. Hydrol.*, *105*(1-2), 157–172, doi:10.1016/0022-1694(89)90101-7.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1-2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K. J. (2009), *Environmental Modelling: An Uncertain Future*?, Routledge, London, U. K.

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

Beven, K. J., and A. M. Binley (2013), GLUE: Twenty years on, *Hydrol. Processes*, *28*, 5897–5918, doi:10.1002/hyp.10082.

Beven, K. J., P. J. Smith, and J. E. Freer (2008), So just why would a modeller choose to be incoherent?, *J. Hydrol.*, *354*(1-4), 15–32, doi:10.1016/j.jhydrol.2008.02.007.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling, *Adv. Water Resour.*, *31*(4), 630–648, doi:10.1016/j.advwatres.2007.12.003.

Bravo, H., F. Jiang, and R. J. Hunt (2002), Using groundwater temperature data to constrain parameter estimation in a groundwater flow model of a wetland system, *Water Resour. Res.*, *38*(8), doi:10.1029/2000WR000172.

Carniato, L., G. Schoups, and N. van de Giessen (2014), Inference of reactive transport model parameters using a Bayesian multivariate approach, *Water Resour. Res.*, *50*, 6406–6427, doi:10.1002/2013WR014156.

Carrera, J., A. Alcolea, A. Medina, J. Hidalgo, and L. J. Slooten (2005), Inverse problem in hydrogeology, *Hydrogeol. J.*, *13*(1), 206–222, doi:10.1007/s10040-004-0404-7.

Cassiani, G., G. Böhm, A. Vesnaver, and R. Nicolich (1998), A geostatistical framework for incorporating seismic tomography auxiliary data into hydraulic conductivity estimation, *J. Hydrol.*, *206*, 58–74.

Christiansen, L., P. J. Binning, D. Rosbjerg, O. B. Andersen, and P. Bauer-Gottwein (2011), Using time-lapse gravity for groundwater model calibration: An application to alluvial aquifer storage, *Water Resour. Res.*, *47*, W06503, doi:10.1029/2010WR009859.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Dahlin, T., and M. Loke (1998), Resolution of 2D Wenner resistivity imaging as assessed by numerical modelling, *J. Appl. Geophys.*, *38*, 237–249.

De Louw, P. G. B., S. Eeman, B. Siemon, B. R. Voortman, J. L. Gunnink, E. S. van Baaren, and G. H. P. Oude Essink (2011), Shallow rainwater lenses in deltaic areas with saline seepage, *Hydrol. Earth Syst. Sci.*, *15*(12), 3659–3678, doi:10.5194/hess-15-3659-2011.

De Louw, P. G. B., S. Eeman, G. H. P. Oude Essink, E. Vermue, and V. E. A. Post (2013), Rainwater lens dynamics and mixing between infiltrating rainwater and upward saline groundwater seepage beneath a tile-drained agricultural field, *J. Hydrol.*, *501*, 133–145, doi:10.1016/j.jhydrol.2013.07.026.

Delsman, J. R., M. J. Waterloo, M. M. A. Groen, J. Groen, and P. J. Stuyfzand (2014a), Investigating summer flow paths in a Dutch agricultural field using high frequency direct measurements, *J. Hydrol.*, *519*, 3069–3085, doi:10.1016/j.jhydrol.2014.10.058.

Delsman, J. R., K. R. M. Hu-a-ng, P. C. Vos, P. G. B. De Louw, G. H. P. Oude Essink, P. J. Stuyfzand, and M. F. P. Bierkens (2014b), Paleo-modeling of coastal saltwater intrusion during the Holocene: An application to the Netherlands, *Hydrol. Earth Syst. Sci.*, *18*, 3891–3905, doi:10.5194/hess-18-3891-2014.

Devos, J., P. A. C. Raats, and R. Feddes (2002), Chloride transport in a recently reclaimed Dutch polder, *J. Hydrol.*, *257*(1-4), 59–77, doi:10.1016/S0022-1694(01)00552-2.

Dobre, S., T. Bastogne, C. Profeta, M. Barberi-Heyob, and A. Richard (2012), Limits of variance-based sensitivity analysis for non-identifiability testing in high dimensional dynamic models, *Automatica*, *48*(11), 2740–2749, doi:10.1016/j.automatica.2012.05.004.

Doherty, J. (2001), Improved calculations for dewatered cells in MODFLOW, *Ground Water*, *39*(6), 863–869.

Doherty, J. (2010), *PEST: Model-Independent Parameter Estimation*, Watermark Numer. Comput., Brisbane, Queensland, Australia.

Doherty, J., R. J. Hunt, and M. J. Tonkin (2011), Approaches to highly parameterized inversion: A guide to using PEST for model-parameter and predictive-uncertainty analysis, *U.S. Geol. Surv. Sci. Invest. Rep.*, *2010–5211*, 71 pp.

Eeman, S., A. Leijnse, P. A. C. Raats, and S. E. A. T. M. Van der Zee (2011), Analysis of the thickness of a fresh water lens and of the transition zone between this lens and upwelling saline water, *Adv. Water Resour.*, *34*(2), 291–302, doi:10.1016/j.advwatres.2010.12.001.

Eeman, S., S. E. A. T. M. Van der Zee, A. Leijnse, P. G. B. De Louw, and C. Maas (2012), Response to recharge variation of thin rainwater lenses and their mixing zone with underlying saline groundwater, *Hydrol. Earth Syst. Sci.*, *16*(10), 3535–3549, doi:10.5194/hess-16-3535-2012.

Feyen, L., K. J. Beven, F. De Smedt, and J. E. Freer (2001), Stochastic capture zone delineation within the generalized likelihood uncertainty estimation methodology: Conditioning on head observations, *Water Resour. Res.*, *37*(3), 625–638, doi:10.1029/2000WR900351.

Feyen, L., J. J. Gómez-Hernández, P. J. Ribeiro, K. J. Beven, and F. De Smedt (2003), A Bayesian approach to stochastic capture zone delineation incorporating tracer arrival times, conductivity measurements, and hydraulic head observations, *Water Resour. Res.*, *39*(5), 1126, doi:10.1029/2002WR001544.

Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross-validation and other methods, *Ground Water*, *45*(5), 627–41, doi:10.1111/j.1745-6584.2007.00341.x.

Foglia, L., M. C. Hill, S. W. Mehl, and P. Burlando (2009), Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function, *Water Resour. Res.*, *45*, W06427, doi:10.1029/2008WR007255.

Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2013), Evaluating model structure adequacy: The case of the Maggia Valley groundwater system, southern Switzerland, *Water Resour. Res.*, *49*, 260–282, doi:10.1029/2011WR011779.

Ginn, T. R., H. Haeri, A. Massoudieh, and L. Foglia (2009), Notes on groundwater age in forward and inverse modeling, *Transp. Porous Media*, *79*(1), 117–134, doi:10.1007/s11242-009-9406-1.

Guggenmos, M., C. Daughney, B. Jackson, and U. Morgenstern (2011), Regional-scale identification of groundwater-surface water interaction using hydrochemistry and multivariate statistical methods, Wairarapa Valley, New Zealand, *Hydrol. Earth Syst. Sci.*, *15*, 3383–3398, doi:10.5194/hess-15-3383-2011.

Gunnink, J. L., J. H. A. Bosch, B. Siemon, B. Roth, and E. Auken (2012), Combining ground-based and airborne EM through Artificial Neural Networks for modelling glacial till under saline groundwater conditions, *Hydrol. Earth Syst. Sci.*, *16*(8), 3061–3074, doi:10.5194/hess-16-3061-2012.

Gusyev, M. A., M. Toews, U. Morgenstern, M. Stewart, P. White, C. Daughney, and J. Hadfield (2013), Calibration of a transient transport model to tritium data in streams and simulation of groundwater ages in the western Lake Taupo catchment, New Zealand, *Hydrol. Earth Syst. Sci.*, *17*(3), 1217–1227, doi:10.5194/hess-17-1217-2013.

Gwo, J., L. Toran, M. Morris, and G. Wilson (1996), Subsurface stormflow modeling with sensitivity analysis using a Latin-Hypercube sampling technique, *Ground Water*, *34*(5), 811–818.
Hassan, A. E., H. M. Bekhit, and J. B. Chapman (2008), Uncertainty assessment of a stochastic groundwater flow model using GLUE analysis, *J. Hydrol.*, *362*(1-2), 89–109, doi:10.1016/j.jhydrol.2008.08.017.
Hendricks Franssen, H.-J., J. J. Gómez-Hernández, and A. Sahuquillo (2003), Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data, *J. Hydrol.*, *281*(4), 281–295, doi:10.1016/S0022-1694(03)00191-4.
Hill, M. C. (2006), The practical use of simplicity in developing ground water models, *Ground Water*, *44*(6), 775–781, doi:10.1111/j.1745-6584.2006.00227.x.
Hill, M. C., and C. R. Tiedeman (2007), *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley, Hoboken, N. J.
Hooghoudt, S. B. (1940), Bijdrage tot de kennis van enige natuurkundige grootheden van den grond - Deel 7. Versl. Landbouwk. Onderz., *46*(14), 515–707.
Hunt, R. J., D. T. Feinstein, C. D. Pint, and M. P. Anderson (2006), The importance of diverse data types to calibrate a watershed model of the Trout Lake Basin, Northern Wisconsin, USA, *J. Hydrol.*, *321*(1-4), 286–296, doi:10.1016/j.jhydrol.2005.08.005.
Hunt, R. J., J. Doherty, and M. J. Tonkin (2007), Are models too simple? Arguments for increased parameterization, *Ground Water*, *45*(3), 254–262, doi:10.1111/j.1745-6584.2007.00316.x.
Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality, *Water Resour. Res.*, *46*, W10517, doi:10.1029/2009WR008584.
Langevin, C. D., D. T. Thorne Jr., A. M. Dausman, M. C. Sukop, and W. Guo (2008), *SEAWAT Version 4: A Computer Program for Simulation of Multi-Species Solute and Heat Transport*, U.S. Geol. Surv., Reston, Va.
Lu, C., P. K. Kitanidis, and J. Luo (2009), Effects of kinetic mass transfer and transient flow conditions on widening mixing zones in coastal aquifers, *Water Resour. Res.*, *45*, W12402, doi:10.1029/2008WR007643.
Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *330*(1-2), 368–381, doi:10.1016/j.jhydrol.2006.04.046.
Molz, F. (2015), Advection, dispersion, and confusion, *Ground Water*, *53*, 348–353, doi:10.1111/gwat.12338.
Nassar, M., and T. Ginn (2014a), Cauchy data requirement of the inverse problem of the mean age equation, *Water Resour. Res.*, *50*, 3583–3588, doi:10.1002/2013WR014674.
Nassar, M., and T. Ginn (2014b), Impact of numerical artifact of the forward model in the inverse solution of density-dependent flow problem, *Water Resour. Res.*, *50*, 6322–6338, doi:10.1002/2013WR014672.
Niswonger, R. G., S. M. Panday, and M. Ibaraki (2005), *MODFLOW-NWT, A Newton Formulation for MODFLOW-2005*, U.S. Geol. Surv., Reston, Va.
Onset (2013), *12-Bit Temperature Smart Sensor (Part #S-TMB-M0XX)*, Bourne, Mass.
Poeter, E. P., and M. C. Hill (1999), UCODE, a computer code for universal inverse modeling, *Comput. Geosci.*, *25*(4), 457–462, doi:10.1016/S0098-3004(98)00149-6.
Post, V. E. A. (2012), Electrical conductivity as a proxy for groundwater density in coastal aquifers, *Ground Water*, *50*(5), 785–792, doi:10.1111/j.1745-6584.2011.00903.x.
Post, V. E. A., H. Plicht, and H. Meijer (2003), The origin of brackish and saline groundwater in the coastal area of the Netherlands, *Neth. J. Geosci.*, *82*(2), 133–147.
Rakovec, O., M. C. Hill, M. P. Clark, A. H. Weerts, A. J. Teuling, and R. Uijlenhoet (2014), Distributed Evaluation of Local Sensitivity Analysis (DELSA), with application to hydrologic models, *Water Resour. Res.*, *50*, 409–426, doi:10.1002/2013WR014063.
Rasa, E., L. Foglia, D. M. Mackay, and K. M. Scow (2013), Effect of different transport observations on inverse modeling results: Case study of a long-term groundwater tracer test monitored at high resolution, *Hydrogeol. J.*, *21*(7), 1539–1554, doi:10.1007/s10040-013-1026-8.
Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Højberg, and L. Troldborg (2012), Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, *Adv. Water Resour.*, *36*, 36–50, doi:10.1016/j.advwatres.2011.04.006.
Risley, J. C., J. Constantz, H. Essaid, and S. Rounds (2010), Effects of upstream dams versus groundwater pumping on stream temperature under varying climate conditions, *Water Resour. Res.*, *46*, W06517, doi:10.1029/2009WR008587.
Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*, W12418, doi:10.1029/2008WR006908.
Rojas, R., L. Feyen, O. Batelaan, and A. Dassargues (2010), On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling, *Water Resour. Res.*, *46*, W08520, doi:10.1029/2009WR008822.
Schlumberger (2010), *Diver Product Manual*, Delft, Netherlands.
Sheets, R. A., M. C. Hill, H. M. Haitjema, A. M. Provost, and J. P. Masterson (2014), Simulation of water-table aquifers using specified saturated thickness, *Ground Water*, *53*(1), 151–157, doi:10.1111/gwat.12164.
Stallman, R. (1965), Steady one-dimensional fluid flow in a semi-infinite porous medium with sinusoidal surface temperature, *J. Geophys. Res.*, *70*(12), 2821–2827.
Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.
Sun, A. Y., R. Green, S. Swenson, and M. Rodell (2012), Toward calibration of regional groundwater models using GRACE data, *J. Hydrol.*, *422-423*, 1–9, doi:10.1016/j.jhydrol.2011.10.025.
Therrien, R., R. G. Mclaren, E. A. Sudicky, and S. M. Panday (2005), *HydroGeoSphere: A Three-Dimensional Numerical Model Describing Fully-integrated Subsurface and Surface Flow and Solute Transport*, Groundwater Simulations Group, Univ. of Waterloo, Waterloo, Ont., Canada, 322 pp.
Thyne, G., C. Güler, and E. P. Poeter (2004), Sequential analysis of hydrochemical data for watershed characterization, *Ground Water*, *42*(5), 711–723.
Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.
Vandenbohede, A., and L. Lebbe (2010), Recharge assessment by means of vertical temperature profiles: Analysis of possible influences, *Hydrol. Sci. J.*, *55*(5), 792–804, doi:10.1080/02626667.2010.490531.
Vandenbohede, A., K. Hinsby, C. Courtens, and L. Lebbe (2011), Flow and transport model of a polder area in the Belgian coastal plain: Example of data integration, *Hydrogeol. J.*, *19*(8), 1599–1615, doi:10.1007/s10040-011-0781-7.
Vandenbohede, A., P. G. B. De Louw, and P. J. Doornenbal (2014), Characterizing preferential groundwater discharge through boils using temperature, *J. Hydrol.*, *510*, 372–384, doi:10.1016/j.jhydrol.2014.01.006.
Van der Meulen, M., et al. (2013), 3D geology in a 2D country: Perspectives for geological surveying in the Netherlands, *Neth. J. Geosci.*, *92*(4), 217–241.

Weerts, H. J. T., W. E. Westerhoff, P. Cleveringa, M. F. P. Bierkens, J. G. Veldkamp, and K. F. Rijsdijk (2005), Quaternary geological mapping of the lowlands of The Netherlands, a 21st century perspective, *Quat. Int.*, *133-134*, 159–178, doi:10.1016/j.quaint.2004.10.011.

Weiss, R., and L. Smith (1998), Efficient and responsible use of prior information in inverse methods, *Ground Water*, *36*(1), 151–163.

Yu, P.-S., T.-C. Yang, and S.-J. Chen (2001), Comparison of uncertainty analysis methods for a distributed rainfall–runoff model, *J. Hydrol.*, *244*(1-2), 43–59, doi:10.1016/S0022-1694(01)00328-6.

Zheng, C. (2009), Recent developments and future directions for MT3DMS and related transport codes, *Ground Water*, *47*(5), 620–625, doi: 10.1111/j.1745-6584.2009.00602.x.

Zhou, H., J. J. Gómez-Hernández, and L. Li (2014), Inverse methods in hydrogeology: Evolution and recent trends, *Adv. Water Resour.*, *63*, 22–37, doi:10.1016/j.advwatres.2013.10.014.