Quantifying Subsurface Parameter and Transport Uncertainty Using Surrogate Modeling and Environmental Tracers

Nicholas E. Thiros[1,*], W. Payton Gardner[1], Marco P. Maneta[1], and Douglas J. Brinkerhoff[2]

[1]Department of Geosciences, University of Montana
[2]Department of Computer Science, University of Montana
[*]Email: nicholas.thiros@umontana.edu

Abstract

We combine physics-based groundwater reactive transport modeling with machine learning techniques to quantify hydrogeologic model and solute transport predictive uncertainties. We train an artificial neural network (ANN) on a dataset of groundwater hydraulic heads and ^{3}H concentrations generated using a high-fidelity groundwater reactive transport model. Using the trained ANN as a surrogate model to reproduce the input-output response of the high-fidelity reactive transport model, we quantify the posterior distributions of hydrogeologic parameters and hydraulic forcing conditions using Markov-chain Monte Carlo (MCMC) calibration against field observations of groundwater hydraulic heads and ^{3}H concentrations. We demonstrate the methodology with a model application that predicts Chlorofluorocarbon-12 (CFC-12) solute transport at a contaminated site in Wyoming, USA. Our results show that including ^{3}H observations in the calibration dataset reduced the uncertainty in the estimated permeability field and infiltration rates, compared to calibration against hydraulic heads alone. However, predictive uncertainty quantification shows that CFC-12 transport predictions conditioned to the parameter posterior distributions cannot reproduce the field measurements. We found that calibrating the model to hydraulic head and ^{3}H observations results in groundwater mean ages that are too large to explain the observed CFC-12 concentrations. The coupling of the physics-based reactive transport model with the machine learning surrogate model allows us to efficiently quantify model parameter and predictive uncertainties, which is typically computationally intractable using reactive transport models alone.

1 Introduction

Predicting the evolution of groundwater quality at contaminated sites is a foremost concern for current and future water resource management. Yet, field-scale contaminant transport that spans decadal to century-long timescales is impractical to directly measure and remains uncertain (Hammond & Lichtner, 2010; Dam et al., 2015; Zachara et al., 2013). It is increasingly apparent that aquifer biogeochemical conditions and contaminant transport dynamics are influenced by groundwater flow with varying residence times (Manning, Mills, Morrison, & Ball, 2015; Bea et al., 2013; Visser, Broers, Van Der Grift, & Bierkens, 2009). For instance, groundwater flow with multi-decadal residence times can impact contaminant reactive chemistry processes (Green, Böhlke, Bekins, & Phillips, 2010; Liao, Green, Bekins, & Böhlke, 2012) and is requisite understanding to estimate transport velocity distributions used to predict contaminant flushing timescales (Manning et al., 2015; Sanford & Pope, 2013; Bohlke & Denver, 1995). Nonetheless, our understanding of groundwater flow and transport at the field-scale is complicated by the limited observations that are sensitive to groundwater with decadal and longer residence times (Zell, Culver, & Sanford, 2018; Gardner, Hammond, & Lichtner, 2015). Further investigation on the role that uncertainties in long-residence time groundwater transport have on field-scale solute transport predictions and predictive uncertainties is needed.

Physics-based numerical models are among the most powerful tools available to assimilate long-residence time groundwater into predictions of subsurface transport (Steefel, DePaolo, & Lichtner, 2005; Li et al., 2017). Given that numerical models are imperfect representations of complex groundwater systems, model calibration using field observations is key to estimate effective model parameters and make solute transport predictions (Doherty, 2015; Hill & Tiedeman, 2007). Model calibration against hydraulic head data alone cannot constrain the groundwater transport velocity fields and leads to poor solute transport predictive performance (Thiros, Gardner, & Kuhlman, 2021; Portniaguine & Solomon, 1998). Augmenting calibration datasets with observations of solute concentrations has been shown to improve estimates of the parameters and processes that control solute transport (e.g. Schilling, Cook, & Brunner, 2019). While injection tracer tests provide solute transport information locally around a well gallery (e.g. Ma et al., 2014), these methods are limited by the time frames of the field campaign and often cannot constrain field-scale processes and parameter heterogeneities. Alternatively, environmental tracers are naturally applied over timescales that range from years to centuries and act as a proxy for solute transport that integrates heterogeneity over broad spatial scales (Cook & Herczeg, 2000; Suckow, 2014). Environmental tracer observations are commonly assimilated into groundwater model calibration datasets (Sanford, Plummer, McAda, Bexfield, & Anderholm, 2004; Portniaguine & Solomon, 1998), with many studies reporting subsequent improvements in hydrogeologic parameter estimates and system hydraulic forecasts (Green et al., 2010; Thiros et al., 2021; Sanford, 2011; Starn, Green, Hinkle, Bagtzoglou, & Stolp, 2014). With respect to field-scale solute transport predictions, Curtis, Davis, and Naftz

(2006) and Åkesson, Bendz, Carlsson, Sparrenbom, and Kreuger (2014) utilize measured tritium (^{3}H) at contaminated field sites to calibrate hydraulic model parameters of reactive transport models that simulate uranium and nitrate, respectively. Despite the prevalence of studies that assimilate environmental tracer observations into model calibration datasets, quantification of model parameter and subsequent solute transport predictive uncertainties has received much less attention.

Theoretical and applied studies have shown that model structural errors, observation data uncertainty, and calibration non-uniqueness degrade groundwater model calibration performance and lead to uncertain predictions (Hill & Tiedeman, 2007; Liu & Gupta, 2007). Calibration non-uniqueness is caused by the inability of field data to constrain the correlations among groundwater flow and transport model parameters and processes (Linde, Ginsbourger, Irving, Nobile, & Doucet, 2017; Doherty & Welter, 2010). Calibration non-uniqueness manifests in many plausible models that can equally fit observation data, thus, interpreting a best-fit model is often fraught and leads to poor predictive performance (Beven, 2006; Hunt, Doherty, & Tonkin, 2007). Quantifying the parameter and resulting model predictive uncertainties, however; remains a challenging task. Typical groundwater model calibration and uncertainty quantification using deterministic methods requires the assumption of model linearization and Gaussian model errors (Tarantola, 2005; Doherty, 2015). However, the non-linearity in the mathematical equations that describe groundwater solute transport and variably saturated flow calls into question the application these simplified uncertainty analysis methods. Studies report parameter uncertainties quantified using using linear approximations can significantly differ from those estimated using more robust Monte Carlo methods (Zell et al., 2018; Yoon, Hart, & McKenna, 2013; Gallagher & Doherty, 2007). Improved uncertainty quantification methods are needed to investigate the impact that including environmental tracer observations that constrain groundwater flow and transport over long temporal scales has on field-scale solute transport predictions and predictive accuracy.

Markov chain Monte Carlo (MCMC) is generally considered the most robust method to perform model calibration and uncertainty analysis (Linde et al., 2017). However, MCMC analysis is computationally intensive and often remains intractable to perform using field-scale groundwater flow and transport models (Yoon et al., 2013; Tonkin & Doherty, 2009). Machine learning based surrogate models that are trained to emulate the input-output response of the high-fidelity groundwater flow and transport model can be used to investigate groundwater system processes and perform uncertainty quantification at a fraction of the computational cost compared to the original physics-based model (Razavi, Tolson, & Burn, 2012; Asher, Croke, Jakeman, & Peeters, 2015). For instance, Laloy and Jacques (2019) and Fienen, Nolan, Kauffman, and Feinstein (2018) compare the ability of multiple surrogate models to emulate physics-based reactive transport models at the column scale and groundwater flow at regional scale, respectively. Recent studies extend the use of groundwater flow and transport surrogate models to perform MCMC analysis to solve the inverse

problem that identifies groundwater contaminant source regions (Zhou & Tartakovsky, 2020; Mo, Zabaras, Shi, & Wu, 2019) and infer subsurface hydraulic conductivity fields (Mo, Zhu, Zabaras, Shi, & Wu, 2019; Rajabi, 2019; Cui et al., 2018; Xu, Valocchi, Ye, & Liang, 2017). To our knowledge, no study has applied surrogate modeling to emulate transport of environmental tracers at the field scales and performed subsequent solute transport predictive uncertainty analysis.

In this work, we develop an artificial neural network (ANN) surrogate model that is trained to simulate groundwater levels and ^{3}H concentrations at a contaminated field site near Riverton, WY. We utilize a high-fidelity, physics-based groundwater flow and transport model to generate the dataset used to train the surrogate model. Using the trained surrogate model as the forward simulator, we perform MCMC calibration to infer uncertainties in subsurface property and hydraulic boundary condition parameters, conditioned on field observations of groundwater levels and ^{3}H observations. Ensembles of groundwater mean age and Chlorofluorocarbon-12 (CFC-12) predictions evaluated using the high-fidelity model and samples from the calibrated parameter posteriors are used to estimate predictive solute transport uncertainties. To investigate the influence that ^{3}H observations that can constrain groundwater transport at the multi-decadal timescales has on solute transport predictive performance, we perform the same MCMC model calibration and predictive analysis using water level observations alone. Through comparison of the model predictive uncertainties given the two calibration datasets, we are able to explore the role that long-residence time groundwater has on solute transport processes at the Riverton site.

2 Methods

2.1 Site Description

The Riverton site is located $\sim$3 km south-west of Riverton, WY on the Wind River Indian Reservation (Figure 1). Contamination at the Riverton site is sourced from a former Uranium and Vanadium processing mill that was active between 1958 and 1963. Despite tailings remediation in 1989 and a risk assessment that predicted natural flushing of the groundwater contaminants would occur within 100 years, elevated Uranium concentrations persist within the shallow alluvial aquifer (DOE, 1998; Dam et al., 2015). A significant amount of groundwater flow and solute transport research has been performed and is on-going at the Riverton site to better understand the Uranium plume dynamics (e.g DOE, 2015; Byrne et al., 2020).

Figure 1:

The Riverton site has an area of $\sim$7 km^2 and is on an alluvial terrace at $\sim$1500 m elevation within the Wind River Basin. The climate is arid to semi-

arid with an annual mean temperature of 8 °C and precipitation of 200 mm that predominantly occurs as winter snow and summer rain (DOE, 2015). The surface hydrology is characterized by the Wind River to the north and Little Wind River to the south. Both the Wind River and Little Wind River experience peak and base flows in June through July and September through February, respectively. The Wind River Basin is composed of interbedded Eocene age sandstone and shale layers (DOE, 1998). Groundwater flows in the southeast direction through three predominant aquifers (DOE, 1998): (1) a 4 to 6 m thick unconfined aquifer comprised of sands, gravel, and silt; (2) a middle semi-confined 5 to 9 m thick sandstone aquifer; and (3) a deep 15 to 20 m thick confined sandstone aquifer. Confining units are composed of shale and have thicknesses up to 10 m.

2.2 Observation Datasets

Field observation datasets are presented in detail within (Thiros et al., 2021). The U.S. Department of Energy (DOE) has performed extensive site characterization at the Riverton site, which includes the installation of numerous groundwater wells and regular groundwater sampling (DOE, 1998, 2015). Throughout this work, observation datasets are from 25 wells finished within the shallow alluvial aquifer. We use a total of 166 water level measurements distributed among the 25 wells for the years 2015 to 2019. The observed water levels are point measurements in time and space and were recorded using a water level tape.

The groundwater ^{3}H and CFC-12 environmental tracer observations are from two separate datasets. Six groundwater ^{3}H observations from 2015 were collected and analyzed by DOE-Legacy Management (DOE, 2015). An additional 22 groundwater ^{3}H, CFC-12, and dissolved noble gas samples were collected in 2019 and 2020 during the months ranging from May to October. Sampling was performed following U.S. Geological Survey procedures (`https://water.usgs.gov/lab`) and chemical analysis was performed at the University of Utah Noble Gas Laboratory following procedures presented in Thiros et al. (2021). CFC-12 concentrations are corrected for excess air calculated using the measured Ne, Ar, Kr, and Xe aqueous noble gas concentrations and the closed-equilibrium excess air model (Aeschbach-Hertig, Peeters, Beyerie, & Kipfer, 1999). Due to expected microbial degradation of CFC-11 and CFC-113, we only use CFC-12 concentrations, which are less likely to be biochemically altered (Cook & Herczeg, 2000).

The atmospheric ^{3}H concentration histories for the Riverton site are inferred from the dataset presented in Michel, Jurgens, and Young (2018) then extended forward in time to 2020 through regression against the Ottawa timeseries. Similarly, atmospheric CFC-12 concentrations are taken from the compiled northern hemisphere timeseries in Bullister (2017) then extended to the year 2020 using measurements made at the NOAA Niwot Ridge Observatory. The atmospheric CFC-12 concentrations are converted to aqueous concentrations in precipitation using Henry's Law at a temperature of 7 °C and elevation of 1502 meters.

2.3 High-Fidelity Forward Model

Transient and 3-D groundwater flow and environmental tracer transport at the Riverton site is simulated using the PFLOTRAN software (Hammond, Lichtner, Lu, & Mills, 2012; Hammond, Lichtner, & Mills, 2014). PFLOTRAN is a physics-based numerical model that solves the fully distributed Richards' equation for subsurface water flow and the advection-dispersion equation for solute transport. Further details on the Riverton site PFLOTRAN model are presented in (Thiros et al., 2021) and briefly described here.

The numerical model domain has an area of approximately 10 km^2 and extends 19 m into the subsurface (Figure 1). Land surface topography of the model was derived from a 1 m resolution digital elevation model (DEM). Model hydrostratigraphic units were simplified to surficial sand and gravel alluvium and underlying sandstone layers (Figure 1). Simulated groundwater levels and environmental tracer concentrations were insignificantly changed when the model included the deeper confined aquifer. The numerical domain is discretized with lateral resolution of nominally 20 m $\times$ 20 m that is further refined to $\sim$5 m approaching the observation well locations. The upper soil and alluvium model layer is discretized into 3 sub-layers that are each 3 m thick and the lower sandstone is a single 10 m thick model layer (Figure 1). The western model boundary approximates a groundwater flowline and is assigned no flow boundary conditions. A no-flow boundary condition is also applied to the base of the model. All other model boundaries (described below) are hydrologically active with boundary condition transience applied at monthly timesteps from the years 1950 to 2020.

The northern and southern model boundaries correspond to the Wind River and Little Wind River, respectively. Hydrostatic boundary conditions that extend from the base of the model to the water surface elevations are applied for the length of both rivers. Transient water surface elevations along the length of the river are extrapolated from the downstream USGS gauging stations using the linear model

$$S(l,t) = R \cdot l + S(l=0,t), \tag{1}$$

where $S(l,t)$ is the estimated river water surface elevation [L] at time t [T] and distance upstream from the USGS gauging station l [L]; R is the water surface elevation slope [L/L]; and $S(l=0,t)$ is the measured water surface elevation at the USGS gauging station [L]. Eq. 1 is solved separately for the Little Wind River (R=lwr) and Wind River (R=wr). While R is varied during the calibration process, the *a priori* value is calculated as the average land surface elevation slopes along the river corridors delineated using the DEM (Table 2). For times that precede the USGS gauging station measurements, $S(l=0,t)$ is approximated as the monthly average of the full measurement records. The eastern model boundary similarly applies a hydrostatic head boundary condition throughout the full depth profile. Here, the water table elevation is linearly interpolated between the estimated Wind River and Little Wind River water surface elevations given by Eq. 1.

The PFLOTRAN model includes groundwater flow and transport through

both the variably saturated vadose zone and fully saturated porous media. Water infiltration into the soil is approximated as temporally variable and spatially homogeneous. The infiltration rates $I(t)$ (L/T) that are applied to land surface using a specified Neumann flux boundary condition are in the form:

$$I(t) = \gamma \cdot I_{th}(t), \tag{2}$$

where γ is a multiplier that scales the base infiltration rate $I_{th}(t)$ [L/T]. $I_{th}(t)$ is calculated as the difference between precipitation rates measured at the Riverton, WY airport (∼10 km away) and evapotranspiration rates evaluated using the Thornthwaite equation (Thornthwaite, 1948). To approximate snowpack processes, we adjust the measured precipitation totals such that precipitation accumulates over days with average temperatures below 0 °C. The accumulated water is then applied to observed precipitation totals for the next day with an average temperature above 0 °C. For the variably saturated flow, the van Genuchten characteristic function (van Genuchten, 1980) is used to relate fluid pressure to effective saturation and the Mualem relation (Mualem, 1976) is used to relate effective saturation to relative permeability. The empirical characteristic function fitting parameters m and α are taken from the literature (Dingman, 2015) and previous modeling studies for the Riverton site (DOE, 1998) (Table 1).

The environmental tracer aqueous concentration histories (described in Section 2.2) are applied to the hydrologically active boundaries using a Dirichlet zero-gradient transport boundary condition. This boundary condition applies a specified concentration (Dirichlet type boundary) for water entering the domain and a zero-gradient Neumann flux condition for water discharging from the domain. We assume that groundwater that enters the sides of the domain deeper than 6 m (two model layers) below the river surface water elevations is pre-modern and does not contain CFC-12 nor ^{3}H. While this assumption is difficult to verify with direct field observations, simulations that applied atmospheric environmental tracer concentrations along the full depth profile below the rivers introduced an unrealistic amount of tracer concentration into the deep model layers.

The porosity (n_{ss}) and permeability (k_{ss}) fields of the sandstone model layer are assumed homogeneous. The hydrogeologic properties of the shallow soil and gravel layers are considered spatially homogeneous for porosity (n_{soil}) and heterogeneous for permeability (k_{soil}). The heterogeneous permeability field is parameterized using the pilot point method, which only varies permeability at discrete locations known as pilot points (Doherty, 2003). To limit the number of parameters, we place 25 pilot points on an unstructured grid with a density that approximates the observation well density (Figure 1). The 3-D $\log_{10}$ permeability field is created by interpolating between pilot points using Ordinary Kriging with an exponential variogram, unit standard deviation for the sill, and isotropic correlation lengths of 800 m in the x and y directions and no variation in the z direction. These correlation lengths were chosen such that all model cells were within approximately 1 correlation lengths of a pilot point. Hydrodynamic dispersion is limited to molecular diffusion as model tests that included

mechanical dispersion had minor impacts on environmental transport compared to the variance in model calibration parameters and boundary conditions.

Fixed parameters used in the high-fidelity PFLOTRAN model (hereinafter referred as high-fidelity model) are summarized in Table 1. Uncertain parameters (described above) that are varied during surrogate model creation and MCMC analysis are collected into the vector $\mathbf{m}$:

$$\mathbf{m} = [n_{soil},\ \log_{10} k_{ss},\ n_{ss},\ \gamma,\ lwr,\ wr,\ \log_{10} k^{p}_{soil}], \tag{3}$$

where $\log_{10} k^{p}_{soil}$ is the permeability of pilot point number $p \in [0, 25]$. Table 2 shows the parameter prior mean values that are based on previous site characterization (DOE, 1998, 2012, 2015). While the high-fidelity model is spatially distributed, we simplify our work flow to only record simulated groundwater levels and ^{3}H concentrations at times and locations that match the observation dataset (Section 2.2). Let $\mathbf{d}^{obs} = [h_{t,x}, {}^3\mathrm{H}_{t,x}]$ be a vector that contains the field observations of groundwater levels h [m] and ^{3}H concentrations [TU] sampled on date t and well location x. Simulated equivalent predictions of $\mathbf{d}^{obs}$ from the high-fidelity model evaluated with parameter vector realization $\mathbf{m}_i$ are represented as $\mathbf{d}^{sim}_i$.

2.4 Neural Network Surrogate Model

A single high-fidelity model run takes up to 10 minutes when distributed over 144 computational cores. Scaling these runtimes to a full MCMC calibration that requires 100000's of model evaluations performed sequentially quickly becomes intractable. To reduce the overall computational expense required by MCMC model calibration and uncertainty analysis, we train an artificial neural network (ANN) surrogate model that approximates the input-output response of the high-fidelity model. Rather than other surrogate modeling techniques (such as Guassian Processes and Polynomial Chaos Expansion), an ANN was chosen due to their flexibility in learning strong model non-linearity and high-dimensional outputs (Asher et al., 2015). While many types of ANN have been effectively trained to emulate hydrologic problems (see. Shen, 2018), they have been used less frequently to emulate groundwater solute transport models at the field scales. In this work, our ANN surrogate model is a deep, but narrow, multi-layer perceptron (MLP). MLP are built as a sequence of layers, each with a number of nodes that are fully connected to all nodes in the previous layer (e.g. Lecun, Bengio, & Hinton, 2015). MLP map from an input parameter layer to an output prediction layer by constructing a series of transformations in the form

$$a_l = ReLU(a_{l-1} W_l + b_l), \tag{4}$$

where a_l is a vector of node activations in the current layer l, a_{l-1} are the node activations from the preceding layer, W_l and b_l are the trainable weight matrices and bias vectors, respectively, and ReLU is the Rectified Linear Unit activation function. The first input layer (a_0) of the MLP has 31 nodes, where each node holds one of the uncertain model parameters in $\mathbf{m}$. The final output

layer has 194 nodes that correspond predictions of water levels (N=166) and ^{3}H concentrations (N=28) at all observation well coordinates and sampling times. Thus, each node in the output layer can be directly compared to its commensurate field observation in $\mathbf{d}^{obs}$ and the high-fidelity model output vector $\mathbf{d}_i^{sim}$ (described above). Between the input and output layers, the MLP has 4 hidden layers, each with 1048 nodes. The structure and hyperparameters of the ANN were determined through manual trail and error tuning. Final water level and ^{3}H predictions from the complete surrogate model evaluated with parameters $\theta = \mathbf{W}, \mathbf{b}$ are notated as $\mathbf{d}_i^{ANN}$.

Training of the MLP refers to tuning the weight and bias parameters in Eq. 4 such that the differences in predictions made by the MLP and the high-fidelity model are minimized. Thus, the data required to train the MLP is generated directly from the high-fidelity model. In particular, we assume the uncertain model parameters in $\mathbf{m}$ come from a uniform distribution with upper and lower bounds given in Table 2. The parameter upper and lower bounds are designed to reflect the the parameter variations measured at the Riverton site, but are enlarged to account for unknown model structural errors and to decrease the likelihood that the trained surrogate model extrapolates during the subsequent MCMC calibration. We then sample 30000 realizations of $\mathbf{m}$ from a quasi-random Sobol sequence that spans the parameter uniform distributions (Sobol, 1998; Brinkerhoff, Aschwanden, & Fahnestock, 2021). Parameter sampling in this way has advantages over random sampling in that the entire parameter space is filled with points that are optimally spread apart, which is critical when a limited number of model runs are being used. High-fidelity model runs that did not converge in a reasonable time were terminated and not included in the training data ensemble. Thus, the final training data ensemble consists of approximately 25000 parameter realizations $\mathbf{m_i}$ and the associated high-fidelity model outputs $\mathbf{d}_i^{sim}$.

The surrogate model is trained to approximate the mapping from $\mathbf{m_i}$ (features) to $\mathbf{d}_i^{sim}$ (predictors) by minimizing the mean-squared error (MSE) loss function

$$I(\theta) = \frac{1}{N}\sum_{i=0}^{N}(\mathbf{d}_i^{sim} - \mathbf{d}_i^{ANN})^2, \tag{5}$$

where N is the number of training examples. Gradients of Eq 5 with respect to θ are calculated using reverse mode automatic differentiation within the Tensorflow python library. Minimization of Eq. 5 and optimization of the MLP parameters θ is then achieved using the ADAM variant of stochastic gradient descent and a batch size of 64. For all MLP model scenarios (described below) we train for a total of 2000 epochs and an initial learning rate of 0.1 that is exponentially decreased every 100 epochs. As a regularization mechanism to prevent overfitting, we apply dropout at a rate of 10% after the hidden layer activation functions in each hidden layer. All ANN input features and output predictors were standardized to have a zero mean and unit variance prior to training.

Evaluating the performance of the MLP requires that the full 25000 member

training data ensemble is split into a subset that is used for training and a subset that is used to validate the MLP predictions. Because the MLP model performance can be a function of how the full dataset is partitioned, we train and validate multiple MLP models using different training-validation splits. In particular, we train 5 MLP networks that have equivalent architectures, but different training and validation sets that are generated using 5-fold cross-validation applied to the full training data ensemble (e.g. Fienen et al., 2018). In doing so, all points within the full training ensemble are seen in both the training and validation sets. The validation root-mean squared errors (RMSE) of the 5 MLP models are averaged to estimate the total MLP accuracy. The final MLP model that is used for subsequent MCMC analysis is trained using the full 25000 training ensemble dataset.

2.5 MCMC Model Calibrations

Bayesian model calibration is a widely used method to infer the posterior distribution of uncertain model parameters $\mathbf{m}$ given observation data $\mathbf{d}^{obs}$ (Linde et al., 2017). The posterior parameter distribution $P(\mathbf{m}|\mathbf{d}^{obs})$, which represents our updated belief in model parameters after considering both observation data and prior knowledge regarding parameters, is quantified using Bayes' theorem

$$P(\mathbf{m}|\mathbf{d}^{obs}) \propto P(\mathbf{d}^{obs}|\mathbf{m})P(\mathbf{m}), \tag{6}$$

where $P(\mathbf{m})$ are the prior parameter distributions and $P(\mathbf{d}^{obs}|\mathbf{m})$ is the likelihood of the observations given a parameters set. Computing the left-hand side of Eq. 6 is intractable and must be approximated using numerical methods. In this work, parameter posterior distributions are quantified using a MCMC method that directly draws discrete samples from the posterior distributions.

The prior distributions reflect the state of knowledge on the parameters before considering the observation data. The *a priori* parameter mean values are established from previous characterization of the Riverton site (Table 2). Following (Brinkerhoff et al., 2021), we define the prior uncertainties using scaled Beta distributions

$$\frac{\mathbf{m} - Bound_L}{Bound_U - Bound_L} \sim \beta(\alpha = 2, \beta = 2), \tag{7}$$

where $Bound_L$ and $Bound_U$ are the lower and upper parameter bounds in Table 2. Beta distribution priors put higher probability density on values in the center of the distribution, which corresponds to the *a priori* parameter mean. However, these priors are vague enough to allow a range of plausible parameters, which is important because unknown model structural errors lead to model calibration parameter estimates that are at incommensurate scales with field-based measurements (Doherty & Welter, 2010). Furthermore, the Beta distributions priors have zero probability density beyond the upper and lower parameter limits. This is advantageous because the surrogate model will not extrapolate to parameter values outside the training ensemble upper and lower bounds.

The likelihood is a measure of goodness of fit between the field observations and an equivalent model prediction. Evaluating the likelihood provides a method to falsify parameter samples drawn from the prior distribution using observation data. Typically, quantifying the likelihood assumes a Gaussian error model (or data noise distribution) that is parameterized using estimates of the assumed observation errors (Linde et al., 2017). In this work there is the added complexity that we are using an imperfect surrogate model to approximate the high-fidelity model (which is also imperfect) during the MCMC analysis. While we do not directly account for the epistemic uncertainty in the high-fidelity model, we incorporate the field observation errors σ_{obs} and surrogate model errors σ_s into a Gaussian likelihood model of the form

$$P(\mathbf{d}^{obs}|\mathbf{m}) \sim N(\mathbf{d}^{obs}, (\sigma_{obs} + \sigma_s)) \tag{8}$$

The field observation errors are treated as a diagonal matrix that includes water level errors set to 0.5 m for all measurements and ^{3}H errors that are 5% of the measured values. Defining the ^{3}H observation errors proportional to the measured concentration prevents larger concentrations from dominating the likelihood function. The surrogate model error is taken as the average RMSE of the 5-fold cross-validation scores obtained during surrogate model training (section 2.4). Augmenting the observation error with the surrogate modeling validation error is a similar approach taken by (Xu et al., 2017). We utilize the computationally-cheap trained surrogate model to evaluate all likelihood function evaluations.

Bayesian inference of model parameters is achieved by sampling from the posterior distribution using the Adaptive Differential Evolution Metropolis algorithm (Ter Braak & Vrugt, 2008) implemented within the pyMC3 python software. This MCMC sampler simulates multiple chains in parallel and uses past states to inform future jumps, which improves the efficiency compared to the Metropolis-Hastings algorithm. We run 10 chains in parallel until each chain contains a total of 100000 samples from the posterior distribution. We evaluate the Gelman-Rubin $\hat{R}$ statistic to ensure the Markov Chains are converged on a posterior distribution (Gelman et al., 2013). We further utilize the Markov Chain traces to qualitatively evaluate an adequate 'burn-in' sample size and exploration of the sample space.

2.6 Posterior Predictive Uncertainties

A goal of this work is to better understand the influence that conditioning model parameters to observations that are sensitive to residence times up to 70 years has on solute transport predictions and predictive uncertainties. To evaluate solute transport predictions at the Riverton site, we randomly sample 1000 instances from the parameter posterior distributions and use the original high-fidelity model to simulate ensembles of the aquifer mean age and CFC-12 concentrations. Mean age is simulated by transporting 'age mass' in the same manner as a conservative solute with a unit strength source term (Goode,

1996). With this formulation of age transport, we do not simulate the full residence time distribution, rather, only the first moment of the distribution. The distribution of mean ages simulated here is not conceptually equivalent to the residence time distribution that is commonly estimated with the use of environmental tracers and lumped parameter models (Cook & Herczeg, 2000; Maloszewski & Zuber, 1982). The predictive mean age distribution represents an estimate in the uncertainty of mean age given plausible parameter sets. This distribution captures the variance in average transport behaviors, conditioned to the observation dataset. Alternatively, residence time distributions are a measure of the flux weighted residences times of the varying flowpaths contributing to a sample, given a single model.

3 Results

3.1 Surrogate Model Performance

The surrogate model is a simplification of the high-fidelity model in that it only predicts groundwater levels and ^{3}H concentrations at locations and times that match the field observation dataset. Figure 2 shows water level predictions made by the trained surrogate model, compared to the validation set, for a subset of 6 observation locations and times (out of 166 total). In particular, the left and right columns contain the three observation locations and times with the lowest and highest water level validation RMSE, respectively. The surrogate model best reproduces the validation water levels at well location 789, with average RMSE and R^2 values of 0.06 m and 1.00, respectively. Alternatively, the highest water level validation inaccuracies occur at well location 700, with average RMSE and R^2 values near 0.35 m and 0.98, respectively. Regressing a line through the surrogate water level predictions at all 166 observation well locations and times in the validation set results in an average RMSE and R^2 of 0.25 m and 0.99, respectively. The slopes of the best-fit lines through the validation water levels all approach unity, suggesting there is little bias in the surrogate model predictions.

Figure 2:

Compared to the water level validation performance, there is considerably more scatter in the ^{3}H concentration validation accuracy (Figure 3). For all 28 ^{3}H observation locations and times, the average RMSE and R^2 are 1.03 TU and 0.87, respectively. The highest and lowest validation ^{3}H concentration RMSE are 0.82 TU and 1.38 TU, which are achieved for wells 853-4 and 722R, respectively. The commensurate R^2 values are 0.84 and 0.78. The slopes of the best-fit lines for all six ^{3}H observations shown in Figure 3 are $\sim$0.85, suggesting a systematic bias in the surrogate model predictions. Comparing the best-fit lines to the one-to-one prediction lines, it is apparent that the surrogate model over-predicts ^{3}H concentrations in the 0 to 3 TU range. Alternatively,

high ^{3}H concentrations (>12 TU) tend to be under-predicted by the surrogate model. This ^{3}H validation error bias correlates with the majority of the ^{3}H concentrations in the full training dataset ensemble being in the ∼4-10 TU range. Thus, there are less training examples in low and high regions of the ^{3}H training ensemble distribution, which likely limits performance in these regions. Tests that expanded the bounds of the prior model did not lead to systematically lower or higher ^{3}H concentrations and often resulted problematic model convergence.

Figure 3:

The water level and ^{3}H MLP training 5-fold cross-validation RMSE are used as a measure of additional model error incurred by using the surrogate model rather than the high-fidelity model in the calibration process (Section 2.5). Including the surrogate model error in the likelihood function (Eq. 8) directly increases the posterior parameter variances, which is essential to not produce overly confident and biased uncertainty estimates for cases when the surrogate model inaccurately emulates the high-fidelity model. The surrogate model error is less than 0.25 m for all water level observation locations and times, which is on a commensurate scale to the assumed measurement errors. The ^{3}H concentration error introduced by the surrogate model ranges from 14% to 88% of the respective field observation and has mean of 32%. For the two ^{3}H observations with measured concentrations that are near 0.5 TU, the surrogate model error approaches 150% of the observed values.

3.2 Parameter Posteriors

Figure 4:

Using the trained surrogate model as the forward simulator, we test two separate model calibration scenarios. The first scenario uses only observed water levels in the calibration dataset. The second model calibration utilizes both observed water levels and ^{3}H concentrations and was performed to further investigate the influence of conditioning solute transport predictions and predictive uncertainty to observations that are sensitive to long-residence time groundwater. In particular, the difference in parameter and predictive uncertainties between the two calibration scenarios represents a measure of observation data worth and can be used to gain insight on the role of long-residence time groundwater in solute transport uncertainties at the Riverton site (Zell et al., 2018). Model calibration and predictive uncertainty quantification methods are consistent between the separate ^{3}H and water level dataset scenarios. The prior and marginal posterior distributions for select model parameters are shown in Figure 4 and the joint posterior distributions are shown in Figure 5 and Figure 6 for the calibration datasets that assimilate ^{3}H and water levels only, respectively.

Model calibration results for the remaining parameters will be discussed below and shown in SI Figure 3.

Figure 5:

Figure 6:

Comparing the prior and marginal posterior distributions provides insight on how much the observation dataset constrains the parameter. For both the sandstone permeability and porosity parameters, there is little difference between the prior and posterior distributions. This indicates that the sandstone layer within the high-fidelity model has minimal impact on the simulated water levels nor ^{3}H concentrations. The sandstone parameter insensitivity is due to the model simulating much greater fluxes in the lateral directions compared to vertical directions in the lower sand and gravel and sandstone layers, which are not significantly impacted by infiltration gradients present at land surface.

For both calibration scenarios shown in Figure 4, the soil porosity posterior distribution shows a slight increase in the mean value and minor uncertainty differences compared to the prior. In particular, the maximum a posteriori is increased to $\sim$35% relative to the prior of 30%, and the uncertainty spans the full bounds of the prior. For the calibration scenario that only uses water level observations, Figure 6 indicates that the soil porosity does not have correlations with the other parameters. Alternatively, the soil porosity has a positive correlation structure with the infiltration rate parameter when the calibration assimilates both water level and 3H observations (Figure 5).

Figure 4 indicates that the infiltration rate parameter (γ) posterior distribution is considerably altered compared to the prior when assimilating ^{3}H observations into the calibration dataset. An infiltration rate parameter of unity corresponds to the difference between measured precipitation and estimated evapotranspiration (Eq. 2). The infiltration rate maximum a posteriori is 0.12, suggesting an infiltration reduction of 88% from the *a priori* estimate. The uncertainty in the infiltration posterior parameter ranges from 0.03 to 0.21, which is significantly reduced compared to the prior distribution. Alternatively, Figure 4 shows that calibration to water levels alone does not lead to comparatively large uncertainty reductions in the infiltration rate posterior. In particular, the infiltration rate posterior distribution when only considering water level observations has a maximum a posteriori of 0.51 and uncertainty regions that closely align with the prior distribution, indicating the model calibration did little to constrain the parameter.

In addition to the infiltration rate multiplier parameter, the gradient in the Little Wind River (lwr) also shows high post-calibration uncertainty reductions when calibrating to the ^{3}H observations. The maximum a posteriori value is increased to 1.44 [m/km] relative to the DEM-based estimate of 1.10 [m/km]. The uncertainty in this increase for the Little Wind River gradient is small relative

to the prior, with lower and upper bounds at 1.40 and 1.50 [m/km], respectively. Similarly, the maximum a posteriori for the Wind River (wr) gradient parameter is 3.05 [m/km], which is increased relative to the prior of 2.59 [m/km]. The uncertainty ranges from lower to upper bounds of 2.57 and 3.57, respectively. The joint posterior distributions (Figure 5) between the Little Wind River and Wind River gradient parameters do not show a significant correlation structure. The Little Wind River gradient has minor positive correlation with the infiltration rate multiplier, which is not the case for the Wind River gradient. This is likely the result of the proximity of numerous observation wells to the Little Wind River. The calibration to water levels alone results in river gradient posterior distributions that closely match those of the ^{3}H calibration scenario (Figure 4). This suggests that the water levels, rather than 3H observations, provide the bulk of the information content that constrains these parameters. However, it is apparent that the Little Wind River gradient parameter has no correlations with the other calibrated parameters when considering only water level observations.

Figure 7:

The permeability pilot point parameters shown in Figure 4 are located near the majority of the observation well locations (see SI Figure 1 for pilot point numbering) and show posterior distributions that are emblematic of the remaining pilot points. Figure 7 depicts the kriged permeability field using the maximum a posteriori $\log_{10}$ pilot point estimates after assimilation of both water levels and ^{3}H observations into the calibration. The kriged field was produced using the same variogram parameters used during the model calibrations. This kriged map corresponds to best estimate of the permeability field after considering both the prior distributions and field observation data. Across all 25 pilot points, the $\log_{10}$ permeability range from -12.6 to -9.6 [m^2]. Despite not providing *a priori* spatial correlation structure within the pilot point prior distributions, the calibration process identifies broad regions of high and low permeability. In particular, there is a cluster of pilot point with permeabilities $\sim 10^{-12}$ [m^2] in the western region of the observation well field. To the north and east of the low permeability zone, the permeabilities approach $\sim 10^{-10}$ [m^2]. Similarly, high permeabilites are predicted for the pilot points near the Little Wind River.

In Figure 7, the size of the plotted pilot point is proportional to the standard deviation of the $\log_{10}$ posterior distribution, which we use as an approximate measure of the post-calibration parameter uncertainty. Comparing the pilot point $\log_{10}$ permeability uncertainties with the prior uncertainty of 1.34 m^2 indicates that all pilot points experience uncertainty reduction due to the calibration process. Generally, pilot points with the least uncertainty reductions are located in the north-east portion of the model domain and do not contain observation wells within a correlation length. Conversely, the pilot points with the lowest $\log_{10}$ posterior uncertainties, which reach $\sim$0.6 m^2, are located near

the observation wells. However, it is apparent that these broad generalizations of the spatial pilot point permeability uncertainties are not consistent throughout the whole domain. For instance, pilot point 21 in the south-east of the model domain is within a correlation length of multiple wells, yet, has one of the largest posterior uncertainties (Figure 7). Comparison of the permeability posterior estimates for the two calibration datasets that include and omit ^{3}H observations shows the pilot point parameter maximum a posteriori estimates are within approximately 1 order of magnitude. The largest discrepancies between the two datasets occur for pilot points 13, 15, and 22, which are shown in Figure 4. It is also apparent that while the pilot point permeability maximum a posteriori estimates can show significant differences between the two calibration dataset scenarios, the posterior uncertainty ranges tend to agree (Figures 4 and SI3).

3.3 Calibration Performance

Figure 8:

The parameter posteriors reflect our prior knowledge regarding parameters and the fit between model predictions and observed data, as shown in Figure 8. In Figure 8(A), the ^{3}H residuals evaluated using the maximum a posteriori parameter estimates (black dots) show considerable scatter and a systematic bias around the one-to-one line. The model calibrated to ^{3}H observations generally over-predicts observations that are below 2 TU and under-predicts observations that are above 4 TU. The uncertainty regions around the maximum a posteriori predictions captures the one-to-one line, however; this uncertainty is commensurate to the variance within the observation dataset. The uncertainties reflect the large ^{3}H surrogate model errors included into the likelihood function. The calibrated surrogate model accurately reproduces the observed water levels, where all of the simulation uncertainties capture the one-to-one line (Figure 8(B)). There are minor observable differences in the water level residuals when ^{3}H observations are omitted from the calibration dataset (Figure 8(D)). Alternatively, Figure 8(C) shows that the calibration to water levels alone significantly over-predicts the ^{3}H observations.

3.4 Predictive Distributions

Figure 9:

We use the high-fidelity model and 1000 parameter sets randomly sampled from the posterior distributions to simulate mean age distributions at four well locations (Figure 9). For the water level and ^{3}H calibration dataset, the solid lines in Figure 9 illustrates that the mean age distributions range from $\sim$0 to

$\sim$400 years. The median of the mean age distribution for well 855-4 is the youngest at $\sim$75 years, while those of wells 859-4 and 860-4 are the highest at sim150 years. Wells 855-4 and 857-4 contain a large fraction of samples with mean ages below 50 years and relatively few samples above 200 years. Alternatively, wells 859-4 and 560-4 have distributions shifted towards older mean ages and contain relatively few samples below 50 years. Using the standard deviation of the mean age prediction ensembles as a proxy for predictive uncertainty, well 855-4 has lowest uncertainty at 53 years and well 859-4 has the highest at 70 years.

The dashed lines in Figure 9 shows the predictive mean age distributions for the surrogate model calibration conditioned to water level observations alone. Compared to including ^{3}H in the calibration dataset, the mean age distributions are similar for well 855-4 where the median of the distribution is 50 years and there is an approximate exponential decay in the amount of samples with older water. Similarly, the mean age distributions at well 857-4 have common shapes, but calibration against ^{3}H shifts the median of the distributions $\sim$ 50 years older. The mean age distributions for wells 859-4 and 860-4 have considerably different shapes in the ^{3}H and water level calibration scenarios. Calibration to water levels alone results in a narrow distribution with a mean of $\sim$ 75 years, while including ^{3}H into the calibration leads to broad distributions with means near 150 years.

Figure 10:

To further investigate the predictive mean age distributions, we compare the ^{3}H and CFC-12 field observations against idealized aquifer mixing behaviors. Figure 10 shows the ^{3}H and CFC-12 concentrations expected in samples assuming no mixing between flow lines (piston-flow model) and mixing that results in an exponential age distribution (Cook & Herczeg, 2000). For both mixing models, the mean ages τ are referenced against the sampling year of 2019. Comparison of the field observations with the mixing models suggests that both the piston-flow and exponential residence time distribution can explain the majority of the observations that are >2 TU. For the piston-flow model, mean ages range from $\sim$30 to 45 years, while the exponential model suggests mean ages ranging from $\sim$30 to 100 years. Figure 10 also suggests that neither the piston-flow nor exponential models can explain the ^{3}H observations that are below $\sim$1 TU.

Figure 11:

We utilize the CFC-12 observation dataset to validate the predictive capability of our calibrated model. We use CFC-12, rather than Uranium, as the predictor of interest because we can apply the typically valid assumption of conservative transport and we know an approximate input concentration function. Figure 11(A) shows predictions of CFC-12 concentrations made using the high-fidelity model and the same 1000 parameter samples used to generate the mean

age ensemble when calibrating to both ^{3}H concentrations and water levels. The CFC-12 prediction ensemble means and predictions using the maximum a posteriori parameters are significantly less then the field observations. In particular, the calibrated model predicts CFC-12 concentrations that are generally <0.5 (pmol/kg), while the field observations are between 1 and 2 (pmol/kg). The lack of high CFC-12 concentration model predictions is consistent with Figure 9 that shows mean age predictions are generally pre-modern (before the year ∼1960). Furthermore, the predictions of CFC-12 concentrations >0 (pmol/kg) suggests the full age distributions contain a fraction of modern water mixing with pre-modern water. It is also apparent that for all but 3 wells, the predictive uncertainty does not capture the field observations.

Figure 11(B) shows the CFC-12 concentration predictions and predictive uncertainties after calibration to water levels alone. Compared to Figure 11(A), the prediction means are larger and show greater variance when ^{3}H is not utilized in the calibration process. Sampling from the parameter posteriors conditioned to water levels alone results in CFC-12 concentration mean predictions that are generally near 1 (pmol/kg). The uncertainties in these predictions spans up to 100% of the mean in most cases. As a result of the higher mean predictions and increased variance, calibration to water levels alone captures the majority of the observed CFC-12 measurements compared to the ^{3}H calibration.

4 Discussion

4.1 Surrogate Modeling Error

Coupling physics-based hydrologic modeling with machine-learning surrogate models to investigate system processes and perform uncertainty quantification is a current area of considerable research (Linde et al., 2017; Shen, 2018; Asher et al., 2015; Razavi et al., 2012). Here, we use an ANN surrogate model trained to emulate a high-performance reactive transport model to investigate the influence that groundwater flow with long residence times has on solute transport predictive uncertainties in a shallow alluvial aquifer. Unlike methodologies that solely use physics-based models for prediction, the success of our uncertainty quantification approach is additionally influenced by the uncertainty and accuracy of the surrogate model.

It is apparent from Figures 2 and 3 that the trained surrogate model has greater ability in learning groundwater level responses compared to ^{3}H concentrations. This result is consistent with studies that show hydraulic pressure fields are more diffuse compared to solute concentration fields in heterogeneous media (Thiros et al., 2021; Voss, 2011). Solute transport has greater sensitivity to the heterogeneity within the permeability field, which leads to a complex mapping from model parameters to ^{3}H concentrations that the surrogate model must learn. While tests showed that the surrogate model training and validation performance was improved when the soil permeability was parameterized with a single homogeneous value, the calibrated model poorly matched the ^{3}H

observation dataset. Degraded surrogate model performance when moving to a heterogeneous field with a larger number of parameters was also reported in (Rajabi, 2019). While we expect that the surrogate model validation accuracy would increase with a larger training set, this was not implemented due to the computational requirements of running the high-fidelity model. However, we assume that the error introduced by the surrogate model is less than the total model error that accounts for the unknown model structural defects (Doherty & Welter, 2010; Xu & Valocchi, 2015).

4.2 System Characterization

The result that the estimated infiltration rates and uncertainties differ whether or not ^{3}H is included within the observation dataset represents the insensitivity of water levels to the infiltration rate parameter. In particular, water level observations cannot uniquely constrain both permeability and recharge parameters (Portniaguine & Solomon, 1998). Alternatively, the reduction of the infiltration rate posterior uncertainties suggests that ^{3}H concentrations in the shallow aquifer are highly sensitive to the boundary condition flux applied at land surface. This sensitivity helps to explain the seemly discrepant result that studies such as Starn et al. (2014) and Carroll, Manning, Niswonger, Marchetti, and Williams (2020) find calibration to environmental tracers significantly constrains porosity estimates, yet, porosity in this work had little influence on the calibration. This is due to the transport velocity field in the shallow subsurface being set by the infiltration rate and the variation in porosity has little influence.

Despite anticipated benefits of calibrating against ^{3}H observations that are sensitive to both the permeability field and infiltration rate, the model poorly predicts CFC-12 concentrations (Figure 11). Comparison of the predictive mean age ensembles for the two calibration dataset scenarios can be used to provide insight on the processes leading to the observed CFC-12 prediction bias. The predictive mean age distribution for the ^{3}H calibration scenario contains a higher proportion of older water relative to the calibration that solely uses water levels. The mean ages that are greater than $\sim$70 years can only explain the observed ^{3}H through a mixture of modern ($<$ 70 years) and premodern ($>$ 70 years) water (Gardner, Susong, Solomon, & Heasler, 2011). In the ^{3}H calibration scenario, the mean ages often exceed 100 years, which suggests the residence time distribution contains a significant fraction of pre-modern water. While constraining the velocity and subsequent groundwater mixing such that there is a large portion of pre-modern water leads to model predictions that can explain the observed ^{3}H concentrations, predictions of CFC-12 are systematically low (Figure 11). This CFC-12 bias is indicative that the fraction of modern, CFC-12 bearing groundwater is too little. This result is unexpected in that both ^{3}H and CFC-12 are sensitive to groundwater ages up to 70 years old.

Despite the contribution of pre-modern groundwater predicted by the numerical model, the environmental tracer mixing plots suggest that the majority of samples can be explained with a simple piston-flow model (Figure 10). The piston-flow model results in single flow paths that do not experience mixing, thus

does not predict a pre-modern component in the samples. The discrepancy in the degree of flowpath mixing between the physics and simple lumped parameter models suggests that the high-fidelity model does not adequately simulate simple flow-lines that experience minimal mixing. As a result, we anticipate that the numerical model must mix a wide range or flowpaths with varying ages to reconcile the ^{3}H observation dataset. A potential reason for this inability of the numerical model to simulate piston-flow scenarios is that the model discretization that is too coarse. A numerical model with large grid cells cannot resolve isolated flow lines that the piston-flow model suggests, which has been previously shown for ^{3}H calibrations by Knowling, White, Moore, Rakowski, and Hayley (2020).

Model structural errors are the result of simplifying and misrepresenting complex systems (Liu & Gupta, 2007). It has been shown that model structural errors, in addition to parametric uncertainties, influence model calibration against environmental tracer information and resulting predictive performance (Thiros et al., 2021). In particular, calibrated model parameters compensate for the ubiquitous, yet unknown model structural errors. Thus, the model parameter combinations that explain the observed ^{3}H concentrations reflect effective parameters that do not accurately generalize to predict CFC-12 concentrations. We anticipate that key model limitations in this work are the assumed structure of the boundary conditions and subsurface heterogeneity. Infiltration is expected to vary spatially and is a function of the complex land surface energy-balance and plant transpiration dynamics that govern evapotranspiration. Our approach that scales a base infiltration rate timeseries by a multiplier does not account for the uncertainty in the temporal dynamics of infiltration. Given the observed sensitivity of ^{3}H simulations to the infiltration rate, this likely imposes a major assumption that propagates to the calibrated parameters and predictive uncertainty.

The simplified subsurface lithology represents an important model assumption that is expected to be influencing the transport simulations. It is apparent that the calibrated model does not accurately capture the full variance of the ^{3}H observations (Figure 8). Similar to findings of Starn and Belitz (2018), this can be explained by the observed ^3H spatial patterns being sensitive to permeability heterogeneity below what was captured with our pilot point spacing. While increasing the number of pilot points could benefit fitting the low and high ^{3}H observations, training the surrogate model would require a larger dataset and computational effort. Despite the indication that the model is under-parameterized with respect to ^{3}H, water level observations are accurately matched. This highlights the limitation in using water levels alone to constrain solute transport predictions.

4.3 Uncertainty Quantification

One of our research objectives was to investigate how uncertainties in long-residence time groundwater flow propagate to field-scale solute transport predictions and predictive uncertainties at the Riverton site. While it is typically

assumed that assimilating more observation data into model calibrations improves model performance, we find that calibrating against ^{3}H observations degrades predictive accuracy. This finding is in contrast to our hypothesis that constraining solute transport predictions with observations of solute transport that span long-residence times will force the model to reconcile a broader spectrum of transport processes, leading to improved system characterization and predictive performance. Rather, our results suggest that due to model structural errors, calibrated parameters take on very specific roles that do not necessarily represent the true processes and properties of the system. Consequently, improved model calibration and lower parametric uncertainties does not necessarily translate to improved model predictions. Identifying the different sources of uncertainties that lead to poor predictive performance is a challenging, yet critical task.

The robust uncertainty quantification methodology that we employ is a valuable step in understanding the model inadequacies that lead to inaccurate predictions. Within the MCMC calibration framework, we identify the full parameter sets that are consistent with the field observations and prior parameter knowledge. Thus, our uncertainty quantification methodology provides some level of confidence that the poor predictive performance we observe is not derived from the limiting assumptions that often influence model calibrations; such as model linearization and solely characterizing local minima with a single optimal model. Alternatively, we are able to attribute the discrepancies between model calibration and predictive performance to model structural errors that are not compensated within the estimated parametric uncertainties. This highlights challenges in assimilating environmental tracer data into complex groundwater models and the need to robustly quantify both parametric and predictive uncertainties when evaluating model performance. However, we also highlight that by assimilating ^{3}H observations and performing uncertainty analysis we are able to expose the presence of model structural errors with high certainty. Alternatively, the water level observations do not contain adequate information content to capture the model defects. These insights support studies that show model structural and conceptual errors can be the dominant source of uncertainty for complex groundwater models (Enemark, Peeters, Mallants, & Batelaan, 2019).

5 Conclusions

We demonstrate a method that facilitates parameter and predictive uncertainty quantification of a computationally expensive physics-based groundwater flow and transport model using a machine learning surrogate model. We train an artificial neural network surrogate model to approximate the groundwater levels and ^{3}H concentrations predicted by a reactive transport model as a function of 31 uncertain model parameters. MCMC analysis is then performed using the trained surrogate model to infer parameter posterior distributions and predictive groundwater mean age and CFC-12 transport uncertainties for a field site near Riverton, WY. We find that while assimilating ^{3}H observations into the

model calibration significantly constrains parameter uncertainties, the model does not predict the observed CFC-12 concentrations. The model parameters and associated uncertainties that explain the ^{3}H observations predict a larger fraction of old groundwater compared to what the CFC-12 observations suggest. The discrepancy between model calibration and predictive performance demonstrates that model misrepresentations can lead to low parametric uncertainties that do not translate to improved model predictive performance nor uncertainty estimates. These finding highlight the need to perform both parametric and predictive uncertainty analysis when assimilating information rich datasets such as environmental tracers into complex groundwater models. The methods presented in this study provide a tool that allows uncertainty quantification using computationally expensive groundwater flow and transport models.

Data Availability Statement

All observation data and supplementary tables and figures are available at `https://zenodo.org/record/5390047#.YTD9xS1h2S4`.

Acknowledgments

This research was funded by Department of Energy NEUP grant number NU-18-MT-UM-040102-04. NET was also supported by National Science Foundation National Research Traineeship DGE-1633831.

The authors thank the Eastern Shoshone and the Northern Arapaho tribes of the Wind River Indian Reservation for site access. We thank Sam Campbell at Navarro Research and Engineering, Inc. for data acquisition and field assistance.

References

Aeschbach-Hertig, W., Peeters, F., Beyerie, U., & Kipfer, R. (1999). Interpretation of dissolved atmospheric noble gases in natural waters. *Water Resources Research*, *35*(9), 2779–2792.

Åkesson, M., Bendz, D., Carlsson, C., Sparrenbom, C. J., & Kreuger, J. (2014). Modelling pesticide transport in a shallow groundwater catchment using tritium and helium-3 data. *Applied Geochemistry*, *50*, 231–239. doi: 10 .1016/j.apgeochem.2014.01.007

Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, *41*, 5957–5973. doi: 10.1002/2015WR016967 .Received

Bea, S. A., Wainwright, H., Spycher, N., Faybishenko, B., Hubbard, S. S., & Denham, M. E. (2013). Identifying key controls on the behavior of an acidic-U(VI) plume in the Savannah River Site using reactive transport modeling. *Journal of Contaminant Hydrology*, *151*, 34–54. doi: 10.1016/ j.jconhyd.2013.04.005

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

Bohlke, J. K., & Denver, J. M. (1995). *Combined use of groundwater dating, chemical, and isotopic analyses to resolve the history and fate of nitrate contamination in two agricultural watersheds, Atlantic coastal plain, Maryland* (Vol. 31; Tech. Rep. No. 9).

Brinkerhoff, D., Aschwanden, A., & Fahnestock, M. (2021, 1). Constraining subglacial processes from surface velocity observations using surrogate-based Bayesian inference. *Journal of Glaciology*, 1–19. doi: 10.1017/ jog.2020.112

Bullister, J. L. (2017). *Atmospheric Histories (1765-2015) for CFC-11, CFC-12, CFC-113, CCl4, SF6 and N2O (NCEI Accession 0164584). NOAA National Centers for Environmental Information. Unpublished Dataset.* (Tech. Rep.).

Byrne, P., Fuller, C. C., Naftz, D. L., Runkel, R. L., Lehto, N. J., & Dam, W. L. (2020). Transport and speciation of uranium in groundwater-surface water systems impacted by legacy milling operations. *Science of the Total Environment*. doi: 10.1016/j.scitotenv.2020.143314

Carroll, R. W. H., Manning, A. H., Niswonger, R., Marchetti, D., & Williams, K. H. (2020, 11). Baseflow Age Distributions and Depth of Active Groundwater Flow in a Snow-Dominated Mountain Headwater Basin. *Water Resources Research*. Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020WR028161` doi: 10.1029/2020WR028161

Cook, P. G., & Herczeg, A. L. (2000). *Environmental Tracers in Subsurface Hydrology*. Springer Science & Business Media.

Cui, T., Peeters, L., Pagendam, D., Pickett, T., Jin, H., Crosbie, R. S., ... Gilfedder, M. (2018, 9). Emulator-enabled approximate Bayesian computation (ABC) and uncertainty analysis for computationally expen-

sive groundwater models. *Journal of Hydrology*, *564*, 191–207. doi: 10.1016/j.jhydrol.2018.07.005

Curtis, G. P., Davis, J. A., & Naftz, D. L. (2006). Simulation of reactive transport of uranium(VI) in groundwater with variable chemical conditions. *Water Resources Research*, *42*(4), 1–15. doi: 10.1029/2005WR003979

Dam, W. L., Campbell, S., Johnson, R. H., Looney, B. B., Denham, M. E., & Steven, C. A. E.-d. (2015). Refining the site conceptual model at a former uranium mill site in Riverton , Wyoming , USA. *Environmental Earth Sciences*, *74*(10), 7255–7265. doi: 10.1007/s12665-015-4706-y

Dingman, S. L. (2015). *Physical Hydrology* (3rd ed.). Long Grove, IL: Waveland Press, Inc.

DOE. (1998). *Final Groundwater Compliance Action Plan (GCAP)* (Tech. Rep.). DOE.

DOE. (2012). *2012 Enhanced Characterization and Monitoring Report Riverton, Wyoming, Processing Site* (Tech. Rep. No. June).

DOE. (2015). 2015 Advanced Site Investigation and Monitoring Report Riverton, Wyoming, Processing Site. *LMS/RVT/S14148*(September).

Doherty, J. (2003). Ground Water Model Calibraton Using Pilot Points and Regularization. *Ground Water*, *41*(2), 170–177.

Doherty, J. (2015). *Calibration and Uncertainty Analysis for Complex Environmental Models*. Brisbane, Australia: Watermark Numerical Computing.

Doherty, J., & Welter, D. (2010). A short exploration of structural noise. *Water Resources Research*, *46*(5), 1–14. doi: 10.1029/2009WR008377

Enemark, T., Peeters, L. J. M., Mallants, D., & Batelaan, O. (2019). Hydrogeological conceptual model building and testing : A review. *Journal of Hydrology*, *569*(July 2018), 310–329. Retrieved from `https://doi.org/10.1016/j.jhydrol.2018.12.007` doi: 10.1016/j.jhydrol.2018.12.007

Fienen, M. N., Nolan, B. T., Kauffman, L. J., & Feinstein, D. T. (2018, 7). Metamodeling for Groundwater Age Forecasting in the Lake Michigan Basin. *Water Resources Research*, *54*(7), 4750–4766. doi: 10.1029/2017WR022387

Gallagher, M., & Doherty, J. (2007, 7). Parameter estimation and uncertainty analysis for a watershed model. *Environmental Modelling and Software*, *22*(7), 1000–1020. doi: 10.1016/j.envsoft.2006.06.007

Gardner, W. P., Hammond, G. E., & Lichtner, P. C. (2015). High Performance Simulation of Environmental Tracers in Heterogeneous Domains. *Groundwater*, *53*(S1), 71–80. doi: 10.1111/gwat.12148

Gardner, W. P., Susong, D. D., Solomon, D. K., & Heasler, H. P. (2011). A multitracer approach for characterizing interactions between shallow groundwater and the hydrothermal system in the Norris Geyser Basin area, Yellowstone National Park. *Geochemistry, Geophysics, Geosystems*, *12*(8), 1–17. doi: 10.1029/2010GC003353

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC. doi: 10.1201/b16018

Goode, D. J. (1996). Direct simulation of groundwater age. *Water Resources Research*, *32*(2), 289–296. doi: 10.1029/95WR03401

Green, C. T., Böhlke, J. K., Bekins, B. A., & Phillips, S. P. (2010). Mixing effects on apparent reaction rates and isotope fractionation during denitrification in a heterogeneous aquifer. *Water Resources Research*, *46*(8), 1–19. doi: 10.1029/2009WR008903

Hammond, G. E., & Lichtner, P. C. (2010). Field-scale model for the natural attenuation of uranium at the Hanford 300 Area using high-performance computing. *Water Resources Research*, *46*(9), 1–31. doi: 10.1029/2009WR008819

Hammond, G. E., Lichtner, P. C., Lu, C., & Mills, R. (2012). CHAPTER 6 PFLOTRAN : Reactive Flow & Transport Code for Use on Laptops to Leadership-Class Supercomputers. *Groundwater Reactive Transport Models*, *5*(2012), 141–159. doi: 10.2174/978160805306311201010141

Hammond, G. E., Lichtner, P. C., & Mills, R. T. (2014). Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLOTRAN. *Water Resources Research*, *50*(1), 208–228. doi: 10.1002/2012WR013483

Hill, M. C., & Tiedeman, C. R. (2007). *Effective Groundwater Model Calibration: With Analysis of Data, Sensitivities, Predictions, and Uncertainty*. Hoboken, New Jersey: John Wiley & Sons, Ltd.

Hunt, R. J., Doherty, J., & Tonkin, M. J. (2007). Are models too simple? Arguments for increased parameterization. *Ground Water*, *45*(3), 254–262. doi: 10.1111/j.1745-6584.2007.00316.x

Knowling, M. J., White, J. T., Moore, C. R., Rakowski, P., & Hayley, K. (2020). On the assimilation of environmental tracer observations for model-based decision support. *Hydrology and Earth System Sciences*, *24*(4), 1677–1689. doi: 10.5194/hess-24-1677-2020

Laloy, E., & Jacques, D. (2019). Emulation of CPU-demanding reactive transport models: a comparison of Gaussian processes, polynomial chaos expansion, and deep neural networks. *Computational Geosciences*, *23*(5), 1193–1215. doi: 10.1007/s10596-019-09875-y

Lecun, Y., Bengio, Y., & Hinton, G. (2015, 5). *Deep learning* (Vol. 521) (No. 7553). Nature Publishing Group. doi: 10.1038/nature14539

Li, L., Maher, K., Navarre-Sitchler, A., Druhan, J., Meile, C., Lawrence, C., ... Beisman, J. (2017). Expanding the role of reactive transport models in critical zone processes. *Earth-Science Reviews*, *165*, 280–301. doi: 10.1016/j.earscirev.2016.09.001

Liao, L., Green, C. T., Bekins, B. A., & Böhlke, J. K. (2012). Factors controlling nitrate fluxes in groundwater in agricultural areas. *Water Resources Research*, *48*(2). doi: 10.1029/2011WR011008

Linde, N., Ginsbourger, D., Irving, J., Nobile, F., & Doucet, A. (2017). On uncertainty quantification in hydrogeology and hydrogeophysics. *Advances in Water Resources*, *110*(May), 166–181. doi: 10.1016/j.advwatres.2017.10.014

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling : Toward an integrated data assimilation framework. *Water Resources Research*, *43*(November 2006), 1–18. doi: 10.1029/2006WR005756

Ma, R., Zheng, C., Liu, C., Greskowiak, J., Prommer, H., & Zachara, J. M. (2014, 2). Assessment of controlling processes for field-scale uranium reactive transport under highly transient flow conditions. *Water Resources Research*, *50*(2), 1006–1024. doi: 10.1002/2013WR013835

Maloszewski, P., & Zuber, A. (1982). Determining the Turnover Time of Groundwater Systems with the Aid of Environmental Tracers 1. Models and Their Applicability. *Journal of Hydrology*, *57*, 207–231.

Manning, A. H., Mills, C. T., Morrison, J. M., & Ball, L. B. (2015). Insights into controls on hexavalent chromium in groundwater provided by environmental tracers, Sacramento Valley, California, USA. *Applied Geochemistry*, *62*, 186–199. Retrieved from `http://dx.doi.org/10.1016/j.apgeochem.2015.05.010` doi: 10.1016/j.apgeochem.2015.05.010

Michel, R. L., Jurgens, B. C., & Young, M. B. (2018). *Tritium Deposition in Precipitation in the United States, 1953–2012* (Tech. Rep.). U.S. Geological Scientific Investigations Report 2018-5086.

Mo, S., Zabaras, N., Shi, X., & Wu, J. (2019, 5). Deep Autoregressive Neural Networks for High-Dimensional Inverse Problems in Groundwater Contaminant Source Identification. *Water Resources Research*, *55*(5), 3856–3881. doi: 10.1029/2018WR024638

Mo, S., Zhu, Y., Zabaras, N., Shi, X., & Wu, J. (2019, 1). Deep Convolutional Encoder-Decoder Networks for Uncertainty Quantification of Dynamic Multiphase Flow in Heterogeneous Media. *Water Resources Research*, *55*(1), 703–728. doi: 10.1029/2018WR023528

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

Portniaguine, O., & Solomon, D. K. (1998). Parameter estimation using groundwater age and head data, Cape Cod, Massachusetts. *Water Resources Research*, *34*(4), 637. doi: 10.1029/97WR03361

Rajabi, M. M. (2019, 2). Review and comparison of two meta-model-based uncertainty propagation analysis methods in groundwater applications: polynomial chaos expansion and Gaussian process emulation. *Stochastic Environmental Research and Risk Assessment*, *33*(2), 607–631. doi: 10.1007/s00477-018-1637-7

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). *Review of surrogate modeling in water resources* (Vol. 48) (No. 7). doi: 10.1029/2011WR011527

Sanford, W. E. (2011). Calibration of models using groundwater age. *Hydrogeology Journal*, *19*(1), 13–16. doi: 10.1007/s10040-010-0637-6

Sanford, W. E., Plummer, L. N., McAda, D. P., Bexfield, L. M., & Anderholm, S. K. (2004). Hydrochemical tracers in the middle Rio Grande Basin, USA: 2. Calibration of a groundwater-flow model. *Hydrogeology Journal*, *12*(4), 389–407. doi: 10.1007/s10040-004-0326-4

Sanford, W. E., & Pope, J. P. (2013, 12). Quantifying groundwater's role in delaying improvements to Chesapeake Bay water quality. *Environmental Science and Technology*, *47*(23), 13330–13338. doi: 10.1021/es401334k

Schilling, O. S., Cook, P. G., & Brunner, P. (2019). Beyond Classical Observations in Hydrogeology : The Advantages of Including Exchange Flux , Temperature , Tracer Concentration , Residence Time , and Soil Moisture Observations in Groundwater Model Calibration. *Reviews of Geophysics*, *57*, 146–182. doi: 10.1029/2018RG000619

Shen, C. (2018, 11). *A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists* (Vol. 54) (No. 11). Blackwell Publishing Ltd. doi: 10.1029/2018WR022643

Sobol, I. M. (1998). On quasi-Monte Carlo integrations. *Mathematics and Computers in Simulation*, *47*(2-5), 103–112. doi: 10.1016/s0378-4754(98)00096-2

Starn, J. J., & Belitz, K. (2018, 9). Regionalization of Groundwater Residence Time Using Metamodeling. *Water Resources Research*, *54*(9), 6357–6373. doi: 10.1029/2017WR021531

Starn, J. J., Green, C. T., Hinkle, S. R., Bagtzoglou, A. C., & Stolp, B. J. (2014). Simulating water-quality trends in public-supply wells in transient flow systems. *Ground water*, *52*, 53–62. doi: 10.1111/gwat.12230

Steefel, C., DePaolo, D., & Lichtner, P. (2005). Reactive transport modeling: An essential tool and a new research approach for the Earth sciences. *Earth and Planetary Science Letters*, *240*(3-4), 539–558. doi: 10.1016/j.epsl.2005.09.017

Suckow, A. (2014). The age of groundwater - Definitions, models and why we do not need this term. *Applied Geochemistry*, *50*, 222–230. Retrieved from `http://dx.doi.org/10.1016/j.apgeochem.2014.04.016` doi: 10.1016/j.apgeochem.2014.04.016

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM.

Ter Braak, C. J., & Vrugt, J. A. (2008, 12). Differential Evolution Markov Chain with snooker updater and fewer chains. *Statistics and Computing*, *18*(4), 435–446. doi: 10.1007/s11222-008-9104-9

Thiros, N. E., Gardner, W. P., & Kuhlman, K. L. (2021, 7). Utilizing Environmental Tracers to Reduce Groundwater Flow and Transport Model Parameter Uncertainties. *Water Resources Research*, *57*(7). Retrieved from `https://onlinelibrary.wiley.com/doi/10.1029/2020WR028235` doi: 10.1029/2020WR028235

Thornthwaite, C. W. (1948). An Approach toward a Rational Classification of Climate. *Source: Geographical Review*, *38*(1), 55–94.

Tonkin, M., & Doherty, J. (2009). Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques. *Water Resources Research*, *45*(1), 1–17. doi: 10.1029/2007WR006678

van Genuchten, M. T. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America*

Journal, *44*, 892–898. Retrieved from https://hwbdocuments.env.nm.gov/LosAlamosNationalLabs/TA54/11569.pdf

Visser, A., Broers, H. P., Van Der Grift, B., & Bierkens, M. F. (2009). Demonstrating trend reversal of groundwater quality in relation to time of recharge determined by3H/3He. *Nederlandse Geografische Studies*(384), 31–46. doi: 10.1016/j.envpol.2007.01.027

Voss, C. I. (2011). Editor's message: Groundwater modeling fantasies-part 2, down to earth. *Hydrogeology Journal*, *19*(8), 1455–1458. doi: 10.1007/s10040-011-0790-6

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

Xu, T., Valocchi, A. J., Ye, M., & Liang, F. (2017, 5). Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. *Water Resources Research*, *53*(5), 4084–4105. doi: 10.1002/2016WR019831

Yoon, H., Hart, D. B., & McKenna, S. A. (2013, 1). Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods. *Water Resour. Res.*, *49*(1), 536–553. doi: 10.1002/wrcr.20064

Zachara, J. M., Long, P. E., Bargar, J., Davis, J. A., Fox, P., Fredrickson, J. K., ... Yabusaki, S. B. (2013). Persistence of uranium groundwater plumes: Contrasting mechanisms at two DOE sites in the groundwater-river interaction zone. *Journal of Contaminant Hydrology*, *147*, 45–72. Retrieved from http://dx.doi.org/10.1016/j.jconhyd.2013.02.001 doi: 10.1016/j.jconhyd.2013.02.001

Zell, W. O., Culver, T. B., & Sanford, W. E. (2018). Prediction uncertainty and data worth assessment for groundwater transport times in an agricultural catchment. *Journal of Hydrology*, *561*, 1019–1036. Retrieved from https://doi.org/10.1016/j.jhydrol.2018.02.006 doi: 10.1016/j.jhydrol.2018.02.006

Zhou, Z., & Tartakovsky, D. M. (2020). Markov chain Monte Carlo with neural network surrogates: application to contaminant source identification. *Stochastic Environmental Research and Risk Assessment*. doi: 10.1007/s00477-020-01888-9

Table 1: Table of PFLOTRAN model parameters.

Parameter	Value	Unit	Description
$^3H_{t_{1/2}}$	12.3287	yr	Tritium half-life
ω	0.39	-	Tortuosity
D_m	10^{-9}	m^2/s	Molecular Diffusion Coefficient
α	10^{-4}	Pa^{-1}	van Genuchten Parameter
m	0.5	-	van Genuchten Parameter

Table 2: Mean, upper, and lower bounds of the uncertain model parameters. $\log_{10}k_i$ refers to the 25 soil pilot point permeabilities.

Parameter	Units	Mean	Lower	Upper	Description
n_{soil}	–	30.0	20.0	40.0	Soil Porosity
$\log_{10}k_{ss}$	m^2	-15.0	-17.0	-13.0	Sandstone Permeability
n_{ss}	–	17.5	10.0	25.0	Sandstone Porosity
γ	–	1.0	0.0	2.0	Infiltration Multiplier
lwr	m/km	1.1	0.6	1.6	Little Wind River Grad.
wr	m/km	2.6	1.3	3.9	Wind River Grad.
$\log_{10}k_i$	m^2	-11.0	-14.0	-8.0	Soil Permeability