2023

Evaluating the Use of Environmental Tracers to Reduce Conceptual Model Uncertainty of Hydrogeologic Models

Andrew Nordberg
University of Montana, Missoula

Jon Graham
University of Montana, Missoula

W. Payton Gardner
University of Montana, Missoula

EVALUATING THE USE OF ENVIRONMENTAL TRACERS TO REDUCE

CONCEPTUAL MODEL UNCERTAINTY OF HYDROGEOLOGIC MODELS

By

ANDREW WILLIAM NORDBERG

Bachelor of Science, New Mexico Tech, Socorro, New Mexico, 2018

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Science
in Geosciences

The University of Montana
Missoula, MT

June 2023

Approved by:

Scott Whittenburg, Dean of The Graduate School
Graduate School

W. Payton Gardner, Committee Chair
Department of Geosciences

Jon Graham
Department of Mathematical Sciences

Marco Maneta
Department of Geosciences

Nordberg, Andrew, M.S., June 2023   Geosciences

Evaluating the Use of Environmental Tracers to Reduce Conceptual Model Uncertainty of Hydrogeologic Models

Chairperson: W. Payton Gardner

Environmental tracer concentrations for CFC12, SF6, and tritium are used in groundwater simulations to assess the ability of these tracers to reduce conceptual model uncertainty due to uncertainty of a site's geologic and recharge characterization. The resulting groundwater simulations are characterized by site-specific hydrologic and geologic data, and with coordination from a field team with years of knowledge about the site. First-order (conceptual) uncertainty is directly addressed by using a stochastic modeling approach for spatial variability of the proposed subsurface configurations. Simulations of environmental tracer concentrations and water levels are used to assess six alternate conceptual models that are based on three alternate geologic interpretations and two levels of spatial complexity in groundwater recharge. Our results show that water levels and tracers both provide unique information, but tracers enhance our ability to distinguish between models throughout multiple analyses. Tracers CFC12 and tritium show how simulating environmental tracer transport in groundwater is better than using water levels at testing alternate hydrogeologic conceptual models and reducing conceptual uncertainty between them.

Table of Contents

1. Introduction

Good quality groundwater ensures global energy, water, food security, and ecological sustainability, and is threatened by anthropogenic forces like over-withdrawal and contamination (e.g. Hu et al., 2010; Troldborg et al., 2010; Johnson et al., 2016). Groundwater studies typically use site-specific, multi-scale hydrologic and geologic data to construct a conceptual-numerical hydrogeologic framework which is then used for building numerical models that simulate groundwater processes and guide resource management decisions (Neuman and Wierenga, 2003). Initial model development steps include: 1) defining the modeling purpose and 2) establishing the conceptual model (Anderson et al., 2015). The purpose describes the problem to solve and selection of appropriate and allowable assumptions or simplifications. The conceptual model is a qualitative and quantitative description or abstraction of the hydrogeologic setting (Neuman and Wierenga, 2003). Enemark et al. (2019) further define the conceptual model as a collection of hypotheses describing the physical structure of the system and mechanistic elements controlling flow and transport (process structure), including a description of parameters (and their spatial variability) used to solve the boundary value problem related to the model's purpose. Subsequent numerical modelling steps include: 3) selection of appropriate equation and computational structures (Gupta et al., 2012), and 4) model testing and evaluation.

The conceptual physical structure includes qualitative and quantitative description of geologic structure, hydrostratigraphic units, lateral extent of the domain, and barriers or conduits (e.g. faults, fractures, channels) that constrain or direct the storage and movement of groundwater or contaminants (Neuman and Wierenga, 2003; Enemark et a., 2019; Gupta et al., 2012; Selroos et al., 2002). The physical structure may also consider aspects such as soil matrix, vegetation, or engineered structures. Many studies use surface geologic mapping (Castro and Goblet, 2003) and topographic profiles to define lateral or vertical extent of the model domain (Troldborg et al., 2007). Well log stratigraphy is used to identify geologic formations and aquifer systems (e.g. Richter, 1981). Borehole cores and geophysical methods help describe subsurface texture distributions in detail (e.g. Dam et al., 2015) and delineate geologic structures at large scales (He et al., 2014). These data can be further interpreted using geospatial techniques including geomorphological facies reconstruction (Houben, 2006), variogram modeling (Ye et al., 2004), indicator geostatistics (Elshall et al., 2013), facies transition probabilities (Carle and Fogg, 1996), multi-point statistics and training images (Remy et al., 2009).

The conceptual process structure is a description of internal and external processes, including physical and chemical reactions, and time-variant boundary conditions related to the purpose of the study (Enemark et al., 2018). Process structure specifies how these dominant processes affect mass and energy fluxes through a groundwater system (Gupta et al., 2012). Many contaminant transport problems may involve processes describing how a contaminant changes phase in subsurface media. Groundwater recharge, defined as the downward flow of water reaching the water table, is the most studied boundary process due to its role in aquifer replenishment (Hu et al., 2019) and establishing pathways for contamination (Dam et al., 2015). A conceptual model of recharge must consider climate, soil, geology, hydrology, surface topography, vegetation, and land use (Healy, 2010).

Uncertainty in groundwater systems is introduced during the initial conceptualization due to spatial-temporal complexity of hydrologic systems and unknown subsurface configuration. This uncertainty can be later amplified if parameter values are calibrated to an incorrect conceptual-numerical framework (Doherty and Welter, 2010). Two model building approaches can be used to deal with conceptual model uncertainty (see Figure 1 of Enemark et al., 2019). The first approach is the consensus approach, where a single model is iteratively developed, tested, and evolves as new data become available. Decreases in uncertainty occur as the model gains complexity in its structures, which effectively turns conceptual (first-order) uncertainty into parameter (low-order) uncertainty when adding more processes to the model's space and time domains (Neuman and Wierenga, 2003). The second approach is the multi-model approach, where an ensemble of alternate models is created that span the plausible conceptualizations that can explain system behavior. These alternate models are differentiated by competing hypotheses, interpretations, levels of complexity in groundwater system structures, or methods to inform model ingredients (Anderson et al. 2015; Enemark et al,. 2019; Neuman and Wierenga, 2003). Including multiple competing conceptual models is believed to aid in uncovering conceptual "surprises" (Bredehoeft, 2005), and is less likely to commit both Type I error (rejection of a valid model) and Type II error (accepting an invalid model) (Neuman and Wierenga, 2003). It is also more likely to directly address first-order (conceptual) uncertainty over low-order (parameter value) uncertainty (Enemark et al., 2019; Suzuki et al., 2008).

Once developed, conceptual models are tested to establish to what degree they are consistent with available data and knowledge (Neuman and Wierenga, 2003; Højberg et al., 2006). Some of the most commonly used model testing datasets are hydraulic head, geophysical data, and hydraulic

conductivity. Bayesian likelihood and least square residuals are favored model-testing and calibration techniques used since they quantify differences between simulated and observed data and allow the modeler to update likely subsurface parameters and confidence in each model. An important consideration throughout the modeling process is that a single dataset cannot be used more than once between developing, calibrating, and testing the models, as this leads to circular reasoning and biased evaluation (Kikuchi et al., 2015; Enemark et al., 2019). Therefore, there is a challenge in choosing a testing dataset which is fit for the modeling purpose, is independent from the development dataset, and has high enough information content to discriminate between models. The most used data source for groundwater model testing is hydraulic head (Enemark et al., 2019). However, this dataset is often not used independent of conceptualization, and may not possess enough information to distinguish between conceptual models.

In this study, we test the hypothesis that including simulated environmental tracer concentrations will result in a decrease in conceptual model uncertainty over the use of hydraulic data alone when testing multiple competing conceptual models. We follow the multi-model approach to develop six reasonable conceptual models based upon the field data and characterization. We then assess environmental tracers' ability to distinguish between and validate these conceptual models with qualitative and quantitative methods to reduce conceptual model uncertainty. This research exemplifies how a relatively common yet uncertain hydrogeologic setting can be tested using practical modeling tools, water levels, and environmental tracers.

2. Theory

2.1 Environmental Tracers

Environmental tracers are chemical species with known atmospheric concentrations from natural and anthropogenic sources. They are applied to global groundwater systems through recharge and/or subsurface radioactive sources and they move in subsurface porous medium influenced by hydrogeologic structures and processes. These include type of stratigraphic layers, their extent, and associated flow and transport properties (hydraulic conductivity, advection, dispersion, diffusion, sorption, and radioactive and biological decay); thereby tracers give information on flow and transport behavior on timescales that range from days to millions of years (Turnadge and Smerdon, 2014). Environmental tracers are typically used to derive apparent groundwater ages or residence time distributions to aid in conceptual model development. Specifically, they have been used to quantify hydrogeologic boundary conditions (Atkinson, et al., 2015; Smerdon

et al., 2014; Smerdon and Gardner, 2021; Rogers, 1958; Genereux and Hemond, 1990; Wanninkof et al., 1990), aquifer connectivity, or recharge in complex subsurface configurations (Batlle-Aguilar et al., 2014); to describe groundwater circulation domains in mountain blocks (Manning and Solomon, 2005), fractured rocks (Maloszewski and Zuber, 1985), and regional hydrothermal systems (Gardner et al., 2011); and identification of contaminant release rates (Chambers et al., 2019; Cook and Herczeg, 2000).

With increasing computing power, Turnadge and Smerdon (2014) suggest numerical groundwater flow and environmental tracer transport simulations (e.g. Gardner, 2015), rather than age estimation, to constrain groundwater flow systems (Thiros et al., 2021). Further, simulating a suite of environmental tracers which covers a wide temporal range can reduce biases or uncover a conceptual surprise from the complex nature of groundwater systems (Cartwright et al., 2017). For example, Troldborg et al. (2007) successfully validated one of four conceptual-numerical models by comparing simulated and observed ^{3}H, ^{3}He, and CFCs concentrations. Ekwurzel et al. (1994) found close agreement between ^{3}H/^{3}He, CFC-11, CFC-12, and ^{85}Kr mean residence time estimations in an alluvial aquifer.

2.2 Geostatistical Modeling

A principal source of uncertainty in groundwater flow and transport modeling is the subsurface structure and parameterization (Neuman and Wierenga, 2003). Literature has shown there are numerous tools to account for uncertainty in the spatial distribution of these structures. Enemark et al., (2019) review how the spatial variability structure of conceptual models is low-order uncertainty and assert that using geostatistical approaches in a stochastic framework allows a modeler to assess the first-order conceptual uncertainty. Spatial statistics methods use techniques like kriging and often use a weighted framework to estimate random variables as a function of the variable's spatial location, with the variable's deterministic (perceived as known) and stochastic (random) characteristics (Cressie, 1991). In this work, there are three determined subsurface structures that are all associated with site characterization data used to condition each model. The unknown geologic spatial heterogeneity of the tested hydrogeologic system is associated with the stochastic part of the hydrogeologic model and is accounted for using geostatistical estimation. We use two structure-imitating methods (Koltermann and Gorelick, 1996) to estimate possible realizations of geologic uncertainty. The first is ordinary kriging.

Two of the conceptual models have uncertainty associated with a clay layer's continuity. Interpreted stratigraphic depth data is assigned as a random variable to be estimated through

ordinary kriging, which uses variogram-based correlation structures. Specifically, ordinary kriging uses a weighted-average framework that includes local variance and spatial covariance information to estimate an unsampled variable in space, $z_{OK}^*(\boldsymbol{u})$:

$$z_{OK}^*(\boldsymbol{u}) = \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha^{OK}(\boldsymbol{u})z(\boldsymbol{u}_\alpha)$$

where $z(\boldsymbol{u}_\alpha)$ is a vector of sample data collected from locations $\boldsymbol{u}_\alpha$, with $n(\boldsymbol{u})$ total locations. Weights, known as kriging weights, λ_α^{OK}, are based on covariances among all points in the sample, and the covariances between sample points and the point to be predicted (Bailey & Gatrell – Ch.5.5). The weights' summation to unity is a constraint accounted for in a system of normal equations with a Lagrange multiplier to minimize estimation error variance. The covariance between the random variables at two locations can be written in terms of the semivariance ($\gamma_{ij} = \mathrm{Var}(Z_i - Z_j)$):

$$C_{ij} = \sigma^2 - \gamma_{ij}$$

where C_{ij} is the covariance between sample points at locations i and j, σ^2 is the population variance of the response values, and γ_{ij} is semivariance. Codes in **R** and SGeMS are used to solve these equations (Pebesma, 2004; Remy, 2011):

$$\begin{cases} \sum_\alpha^n \lambda_\alpha \, \gamma(\boldsymbol{u}_\alpha, \boldsymbol{u}_1) - \varphi & = & \gamma(\boldsymbol{u}_0, \boldsymbol{u}_1) \\ \sum_\alpha^n \lambda_\alpha \, \gamma(\boldsymbol{u}_\alpha, \boldsymbol{u}_2) - \varphi & = & \gamma(\boldsymbol{u}_0, \boldsymbol{u}_2) \\ & \dots & \\ \sum_\alpha^n \lambda_\alpha \, \gamma(\boldsymbol{u}_\alpha, \boldsymbol{u}_n) - \varphi & = & \gamma(\boldsymbol{u}_0, \boldsymbol{u}_n) \\ \sum_a \lambda_\alpha = 1 & & \end{cases}$$

where φ is the Lagrange multiplier, the semivariance $\boldsymbol{\gamma(u_i, u_j) = Var[Z(u_i) - Z(u_j)]} \; \forall i, j$ can be estimated using variogram modeling: experimental directional variograms offer a visual depiction of spatial correlation, including azimuth, correlation length, and anisotropy in cartesian space. Spatial correlation can then be captured by parameterizing theoretical variogram models.

During manual exploratory data analysis in this study, the shape of observed experimental variograms was adequately represented by a spherical model:

$$\gamma(\boldsymbol{h}) = \begin{cases} 0 & h = 0 \\ a + \sigma^2 \left(1.5 * \dfrac{\|\boldsymbol{h}\|}{r} - 0.5 * \left(\dfrac{\|\boldsymbol{h}\|}{r} \right)^3 \right) & 0 < h \leq r \\ a + \sigma^2 & h \geq r \end{cases}$$

where $\boldsymbol{h}$ = the spatial (Euclidean) distance between sample points, a is the nugget of the variogram, σ^2 is the sill, and r is the range. Such variogram models can be examined in specific directions (directional variograms) to assess whether anisotropy is present and to determine the direction of maximum anisotropy under an assumption of geometric anisotropy.

The second method uses Training Images and multiple-point statistics to simulate subsurface geomorphic facies that plausibly facilitate groundwater transport. Verbatim from Remy et al., 2009, a training image (TI) is a representation of how values are jointly distributed in space; it is an unconditional realization of a random function model that is a prior conceptual depiction of the distribution in space of those values; however, they are not yet conditioned to spatial information contained in the data set. In plain language, TIs are a numerical description that can encapsulate geometry and physical structures of complex geologic media. Parameterization of TIs is directly conditioned to features of a geologic system, including feature shape, orientation, sinuosity, and facies proportions. Gridded realizations of TIs are populated by indicators of occurrence of an event:

$$I_k(\boldsymbol{u}) = \begin{cases} 1 & \text{if the event } k \text{ occurs at location } \boldsymbol{u} \\ 0 & \text{if not} \end{cases}$$

where event k could be the presence of a specific geomorphic facies at location $\boldsymbol{u}$.

By extending indicator methods to the kriging paradigm and considering all sample points at once, i.e., in a multiple-point statistical framework (see Remy et al., 2009 for detailed theory), an extended form of the simple indicator kriging estimator is:

$$I_{SK}^*(\boldsymbol{u}) = Prob^* \{I(\boldsymbol{u}) = 1 | n(\boldsymbol{u})\} =$$
$$p_o \quad (\text{prior probablility for } I(\boldsymbol{u}) = 1)$$
$$+ \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha^{(1)}(\boldsymbol{u})[I(\boldsymbol{u}_\alpha) - p_o] \quad (\text{one at a time})$$
$$+ \sum_{\alpha=1}^{(n(\boldsymbol{u}),2)} \lambda_\alpha^{(2)}(\boldsymbol{u})[I(\boldsymbol{u}_{\alpha 1})I(\boldsymbol{u}_{\alpha 2}) - E\{I(\boldsymbol{u}_{\alpha 1})I(\boldsymbol{u}_{\alpha 2})\}] \quad (\text{two at a time})$$
$$+_{...}$$
$$+ \lambda_\alpha^{(n(\boldsymbol{u}))}(\boldsymbol{u}) \left[\prod_\alpha^{n(\boldsymbol{u})} I(\boldsymbol{u}_\alpha) - E\left\{ \prod_\alpha^{n(\boldsymbol{u})} I(\boldsymbol{u}_\alpha) \right\} \right] \quad (\text{taken all together})$$

which simplifies to the Single Normal Equation solution:

$$I_{SK}^*(\boldsymbol{u}) = \lambda \cdot \prod_{\alpha=1}^{n(\boldsymbol{u})} I(\boldsymbol{u}_\alpha)$$

where λ is a single kriging weight. Remy et al., 2009 explain that the exact expression of the conditional probability is identified by this equation by implementing the Bayes relation, and that the individual conditional probabilities can be obtained by the facies distribution information stored in a training image.

Implementation of this theory is accessible by using the tiGenerator utility and Single Normal Equation Simulation (SNESIM) tool found in SGeMS software, whereby the SNESIM algorithm reads the conditional facies distribution from TIs and stores that information in a search tree (Strebelle, 2002). The algorithm conditions that information to hard spatial data, and progresses one pixel at a time on a simulation grid until the realization is complete, fulfilling conditional facies distributions and other user-specified requirements with spatially accurate realizations (Strebelle, 2002; Journel, 2002; Høyer, 2016).

2.3 Groundwater Flow and Solute Transport Modeling

The groundwater flow equation, which describes movement of constant density water in porous earth material under fully saturated, transient, anisotropic, heterogeneous conditions, and considers sources and sinks, is:

$$\frac{\partial}{\partial x}\left(K_{xx}\frac{\partial h}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_{yy}\frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right) + W = S_s\frac{\partial h}{\partial t}$$

where K_{xx}, K_{yy}, K_{zz}, are values of hydraulic conductivity along the x, y, and z Cartesian coordinate axis [L/T], h is hydraulic head [L], W is a volumetric flux per unit volume representing sources (positive W) and sinks (negative W) of water [T^{-1}], S_s is the specific storage of the subsurface [L^{-1}], and t is time [T]. MODFLOW 2005 solves the groundwater flow equation using a finite-difference block centered code, yielding values of head at specified locations and times (Harbaugh, 2005).

MT3DMS is a modular three-dimensional multispecies transport model that simulates advection, dispersion, diffusion, and chemical reactions of solutes in groundwater systems (Zheng and Wang, 1998). The partial differential equation used by MT3DMS is:

$$\frac{\partial(\theta C^k)}{\partial t} = \frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C^k}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(\theta v_i C^k) + q_s C_s^k + \sum R_n$$

where C^k is the dissolved concentration of species k [ML^{-3}], θ is the porosity of the subsurface medium [-], t is time [T], x_i is the distance along the respective Cartesian coordinate axis [L], D_{ij} is the hydrodynamic dispersion coefficient tensor [L^{2}T^{-1}], v_i is the seepage or linear pore water velocity, [LT^{-1}], q_s is the volumetric flow rate per unit volume of aquifer representing fluid sources (positive) and sinks (negative) [T^{-1}], C_s^k is the concentration of the source or sink flux for the species k [ML^{-3}], and $\sum R_n$ is the chemical reaction term [ML^{-3}]. The code simulates changes in concentrations of miscible compounds using seepage velocity fields from a MODFLOW model and various solute boundary conditions.

3. Methods

This project uses a suite of three environmental tracers to test an ensemble of conceptual models for a contaminated site near Riverton, WY. The methods include a description of the site location, including regional geology, stratigraphy and lithology descriptions with aquifer characterization, site-specific sampling, development of alternate conceptual models, stochastic modeling of the alternate physical structures, applying a deterministic proxy model of a process structure, and fully physics-based modeling of groundwater flow and environmental tracer migration using finite difference solvers from MODFLOW5 and MT3DMS.

3.1 Site Description

The study site is in the Wind River Basin of central Wyoming, USA (Figure 1). The Wind River Basin is an asymmetric foreland basin with origins in Pre-Cambrian time (Keefer et al., 1970). Thick clastic and organic sedimentary packages comprise the majority of basin fill, deposited throughout numerous episodes of epicontinental sea level fluctuations, basin subsidence, and orogeny (Richter, 1981). The two principal aquifers under investigation at the site include the Wind River Formation and a surficial alluvial aquifer. The Wind River Formation consists of consolidated sands, gravels, and muds deposited in fan, fluvial, and lacustrine environments following mountain uplift during the Eocene (Sinclair et al., 1911; Richter, 1981). These sediments have complex and discontinuous structures with variable lithologies and thicknesses (30 to >600 meters) depending on basin locality. The surface Quaternary deposits are flood-plain alluvium, terrace, and slope wash deposits, and other unconsolidated gravelly sediments. Alluvium is established throughout the basin, ranging in thickness from 1 meter to more than 30 meters (Daddow, 1996). These deposits are mainly composed of unconsolidated gravel, sand, silt, and cobbles. Throughout much of the basin, the Wind River formation directly underlies flood-plain alluvium sediments and outcrops extensively at a regional scale.

8

The site is located on an alluvial flood-plain terrace located 3.9km southeast of Riverton and is bounded by two rivers, the Wind River 1.6km north and Little Wind River 1.0km south, with their confluence 4km to the east. Both rivers are hydraulically connected to groundwater occurring in the alluvium. Nearby USGS stream stations measure daily discharge and gage height for each river. The topography at field scale is relatively flat with an average gradient of 0.002 dipping to the southeast. Variable-scale surface features including small dry channels, roads, ditches, irrigation channels, and larger paleo river channels, including an oxbow lake, are noted from field visits and aerial imagery. Given the local geologic and climatic history and interpretations in nearby locations, it is theorized that these landscape-scale paleo channels formed during glacial maxima as large fluvial processes cut across the landscape, leaving behind high terraces, gravelly channels, and silty floodplain material.

The climate is arid to semi-arid (DOE, 1998; Richter, 1981) with an average precipitation of 200mm occurring during the stormier months of April through June. Vegetation primarily consists of bunchgrasses, shrubs, and some trees, with generally greener areas concentrated near surface waters. Evapotranspiration is thought to contribute to upward migration of contaminants, is a function of plant types and location, and varies on an annual basis (Dam et al., 2015). Precipitation and stream hydrographs have similar character, indicating that the rivers are primarily fed by precipitation events, although they also may receive contributions from spring snow melt from mountains at the basin margins.
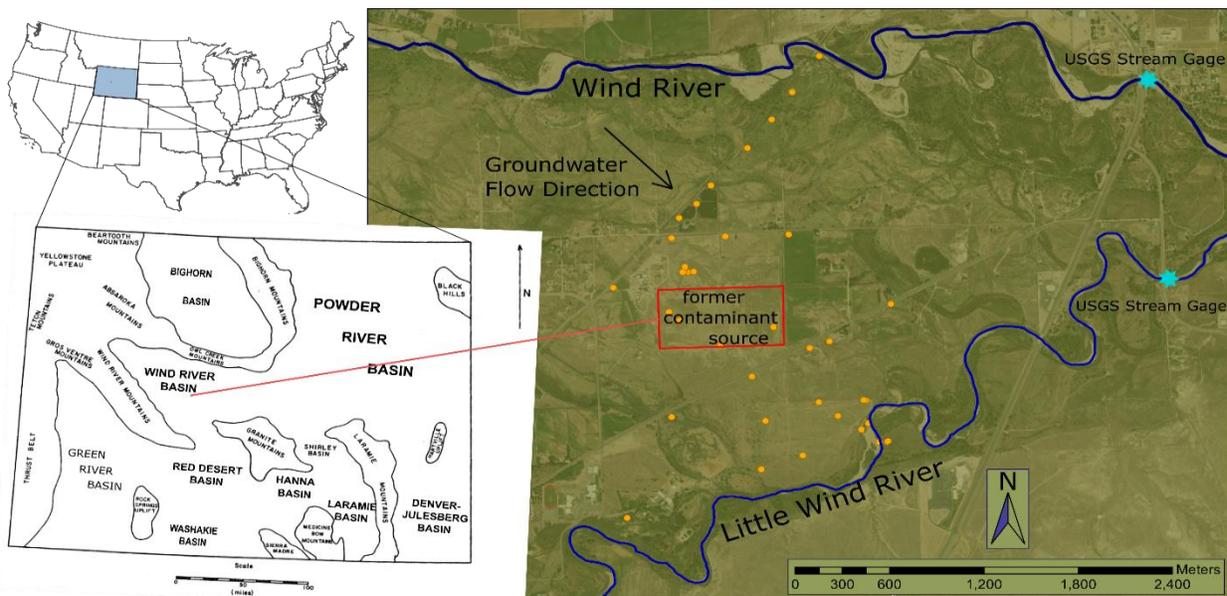
Figure 1: The study site is located at the center of the Wind River Basin, bounded by the Wind River Range to the west, Granite Mountains to the south, and Owl Creek Mountains to the north. The mapped orange points represent U.S. DOE logged monitoring wells throughout the site.

The site itself is at a former uranium and vanadium milling area that operated from 1958 to 1963. Its operations produced slurries of waste which fed into a 29-hectare unlined impoundment. The tailings are thought to be the primary contamination source (White et al., 1984; Narasimhan et al., 1986; DOE 1998) and were relocated to the Gas Hills East disposal site 72 kilometers away (DOE, Fact Sheet). Previous hydrogeologic characterization and modeling followed the consensus approach, which included a single layer aquifer with heterogeneous hydraulic conductivity fields. This conceptual and its associated radionuclide transport model guided a Natural Flushing strategy to comply with the EPA's groundwater quality standards. However, large flooding events in 2010 and 2015 resulted in remobilization and increased concentrations of uranium within the unsaturated zone and surficial aquifer. Ongoing groundwater characterization and modeling by the DOE are being carried out to assess groundwater flow and transport conditions, specifically, an investigation of soil and aquifer geochemical properties that control the fate of these contaminants (Dam et al., 2015). Recent site characterization efforts by the DOE include eight new monitoring wells completed in the shallow aquifer and vadose zone, infiltration and slug tests, additional geochemical and environmental tracer samples, and temporary Geoprobe boreholes to investigate soil properties and hydrodynamic dispersion at specific locations.

The purpose of the modeling exercise in this thesis is to evaluate alternative conceptual models for the Riverton site and whether they may aid in predicting contaminant transport pathways controlled by groundwater flow by reducing uncertainty between models. Specifically, we evaluate uncertainty in physical structure and groundwater recharge to the hydrogeologic system.

The earliest DOE reports highlight three principle hydrogeologic units: an unconfined Quaternary gravel and sand aquifer, underlain by two upper Eocene Wind River units: a leaky shale (clay) aquitard and a semiconfined sandstone aquifer. The conceptual geologic cross section from DOE, 1998 shows a pinching Wind River Sandstone unit contained within the alluvium that ends near the center of the site and otherwise is assumed to be consistent throughout the northwestern domain. Dam et al., 2015 lump this pinching sand into the surficial aquifer and assume relatively constant clay and alluvial aquifer thicknesses. In this study, drill logs from the Department of Energy's Office of Legacy Management are utilized to evaluate new alternate conceptual physical structures for the site. Logs were digitized and interpreted to create geologic cross sections throughout the site. Cross section locations were chosen to adequately delineate the site's stratigraphy (Figure 2, inset).
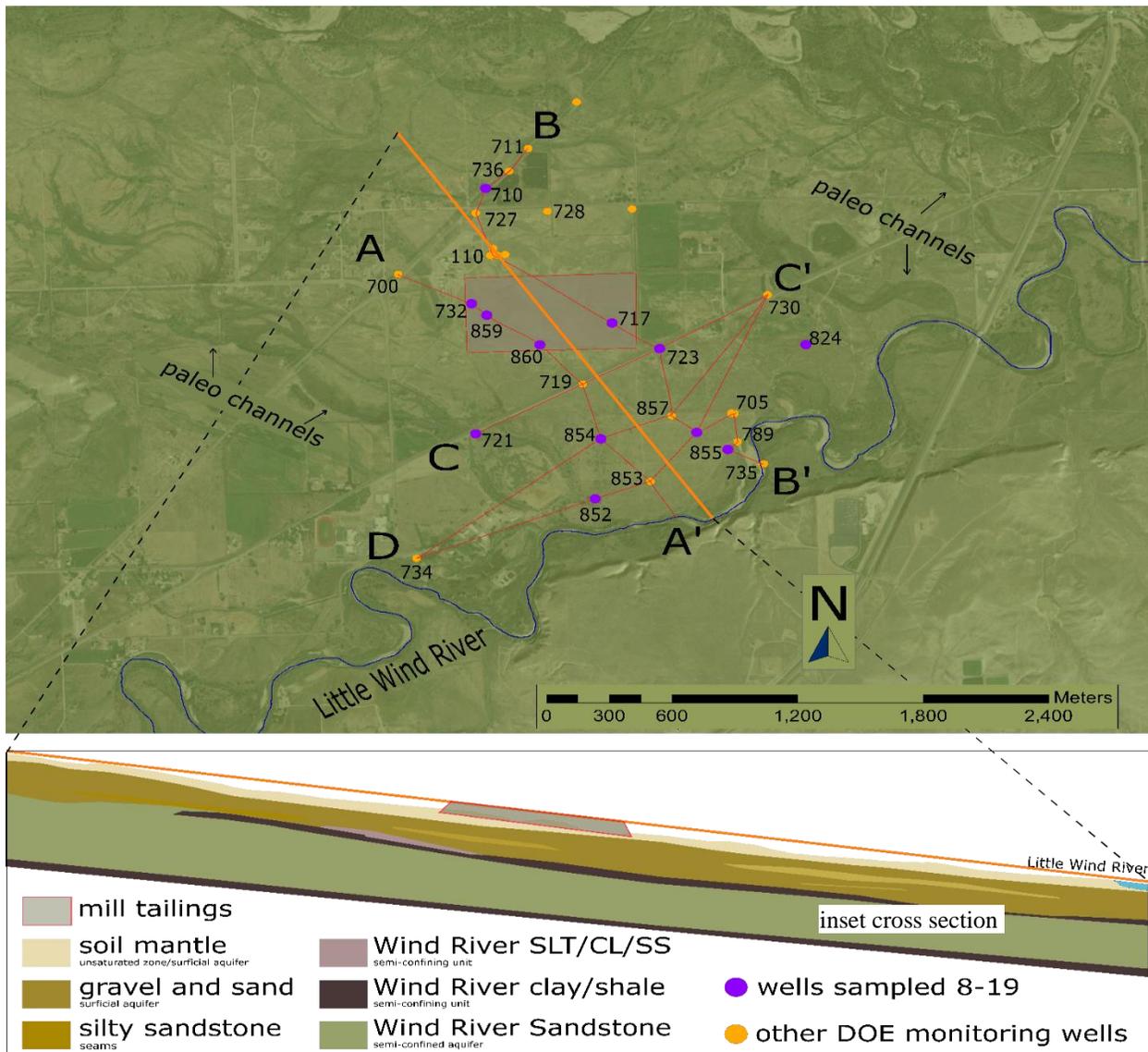
Figure 2: Map view of location of the general stratigraphy cross section (orange line), with additional lettered cross section (A-A', B-B', C-C', and D-D') used to interpolate lithologies and geologic structure (i.e. stratigraphy). Also indicated on the map are surface paleo channel locations. Orange dots are U.S. Department of Energy monitoring well locations used during field activities or in site characterization. Those indicated in purple were sampled during August 2019, have the best coverage along the cross section and are illustrated in the Results section.

Boreholes characteristically show general sedimentary boundaries, which in descending order include surface and subsurface soil with silt and loam, sands and gravels, silty shale and clay, Wind River sandstone, and additional Wind River shale. Uncertainty in well logs' descriptions is noted at multiple well nests. Specific examples include: one well log distinctly marking the clay, with another log contains no clay at the same depth and location; or multiple logs simply showing undifferentiated sand, silt, and clay. Further, some wells are finished seemingly shallow in the surficial gravel, yet describe being in Wind River sandstone; or some wells are only partially through the clay and therefore the extent and complexity of the clay is unknown.

Determination of this clay layer, i.e. physical structure, may be essential in evaluating groundwater mixing between aquifers. If the clay layer is discontinuous, there may be elevated connectivity between the sandstone aquifer with the upper alluvial aquifer layer; mixing by upwelling from the semi-confined sandstone aquifer into the surficial aquifer could occur. This process is hypothesized in this work as contributing to remobilization of uranium during the flooding events. This was hypothesized after seeing higher concentrations of uranium (Goble, 2018) seemed to occur in high uncertainty areas. Specifically, where well logs show there to be no clay, and instead that Wind River Sandstone directly underlays alluvium and is closer to the surface than in other areas. However, following an interpretation of a more continuous-clay layer, and that these wells were not deep enough to reach the clay, both clay and sandstone formations would remain deeper at those locations and would maintain more separation of groundwater between aquifers. In this conceptualization, three or four layers must be considered including vadose zone soil, saturated sand and gravel, a semi-confining clay, and the lowest sand member confined at the bottom with shale. Uncertainty is based on the interpretation of the clay's physical structure, whether the clay is more continuous or discontinuous, which accounts for two possible conceptual physical structures for the Riverton site (Figure 3A) and Figure 3B)).

Channel-like structures are observed from site visits and are seen in aerial photography. The third alternate conceptualization, which is specific to alluvium structure explores how paleo channels could influence groundwater flow and transport at the site by acting as a conduit for faster travel (Figure 3 C)). This hypothesis is supported by USGS aquifer tests recording high hydraulic conductivity in the current Little Wind River channel: 178 meters per day (m/day), which is more than 7x higher than the average background hydraulic conductivity of 22.9 m/day. Variations of models with paleo channels overlying the semiconfined sandstone are not considered.
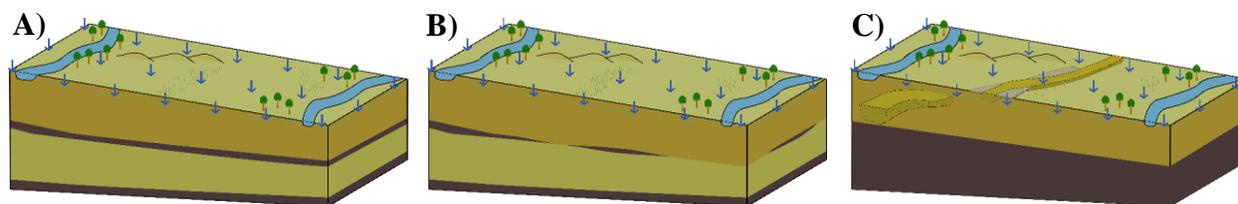
Figure 3: Conceptual physical structure of the hydrogeologic system at Riverton, differentiated by A) continuous-clay conceptual model (cCM), B) discontinuous-clay conceptual model (dCM), and C) paleo channels conceptual model (pcCM). The top recharge boundary assigned to these models assumes a spatially uniform rate that is 25% of the measured precipitation at Riverton.

Evapotranspiration and its effect on groundwater recharge is especially important at sites with a shallow water table, when considering gaseous diffusion in complex multi-phase media (Doble and Crosbie, 2017), or contamination leachability in the solid phase (Dam, 2015), and to avoid

underestimation of groundwater evapotranspiration ([Balugani et al., 2017](#)). In this work, two uncertain evapotranspiration (ET) structures are considered: spatially varying and spatially uniform. The spatially uniform assumption is associated with a simple conceptualization that spatial variability of recharge and infiltration flux does not matter. Effects of ET propagate to estimated recharge (R) fluxes, which is conditioned as a constant percentage of estimated regional precipitation (25%).

The second structure assumes spatial heterogeneity of ET will propagate to R and will impact groundwater flow and transport processes. Primary controls of ET include interception and transpiration, which are directly related to leaf area index (LAI) ([Wang, 2008](#)). To estimate potential effects that spatially varying ET may have on groundwater recharge, remotely sensed vegetation with proxy LAI data and soil characteristics were used in a heuristic model ([Simic et al., 2014](#)). Assuming no overland flow, precipitation is partitioned into R and ET. Three classes of LAI were determined from aerial photography, including background, mid, and high-level LAI thresholds ([Figure 4](#)). Areas depicted in high LAI areas are seen associated with tree cover in channel and floodplain areas with fine-grained silts and clay. Background LAI is associated with the open and dry prairie areas with sand and gravel. The mid-range is associated with bushy, grassy, or some farmland areas with sandy ground. Non-uniform recharge to the water table is then calculated as the ratio between interpreted recharge and evapotranspiration values from Simic et al. (2014) given the LAI conditions:

$$R_{RIV} = \frac{R_{OAR}}{R_{OAR} + ET_{OAR}} * P_{RIV}$$

where R_{RIV} is estimated recharge for Riverton, R_{OAR} and ET_{OAR} are the recharge and evapotranspiration values from Simic et al. (2014) given the specified LAI, and P_{RIV} is the measured precipitation at Riverton. This partitioning methodology, with recharge ratios and percentage of recharge summarized in [Table 1](#), is deterministically applied throughout the Riverton domain as in [Figure 4](#). This hereby describes two conceptual process structures to consider in the tested hydrogeologic system; the three models with alternate physical structures previously described will also be evaluated with both spatially variable and non-uniform recharge applied, as summarized in [Table 1](#), for a total of six conceptual models (CMs).

Table 1: The percentage of total precipitation expected to enter the groundwater system at Riverton using aerial imagery.

LAI proxy value	ET [mm\yr]	Recharge amount	% Precipitation	reference well
Background, uniform	-	-	25	RVT-860
LAI = 2 or 3	260	385	29.8	RVT-727
LAI = 4 or 5	320	280	23.3	RVT-700
LAI = 7 or 8	390	150	13.9	RVT-721
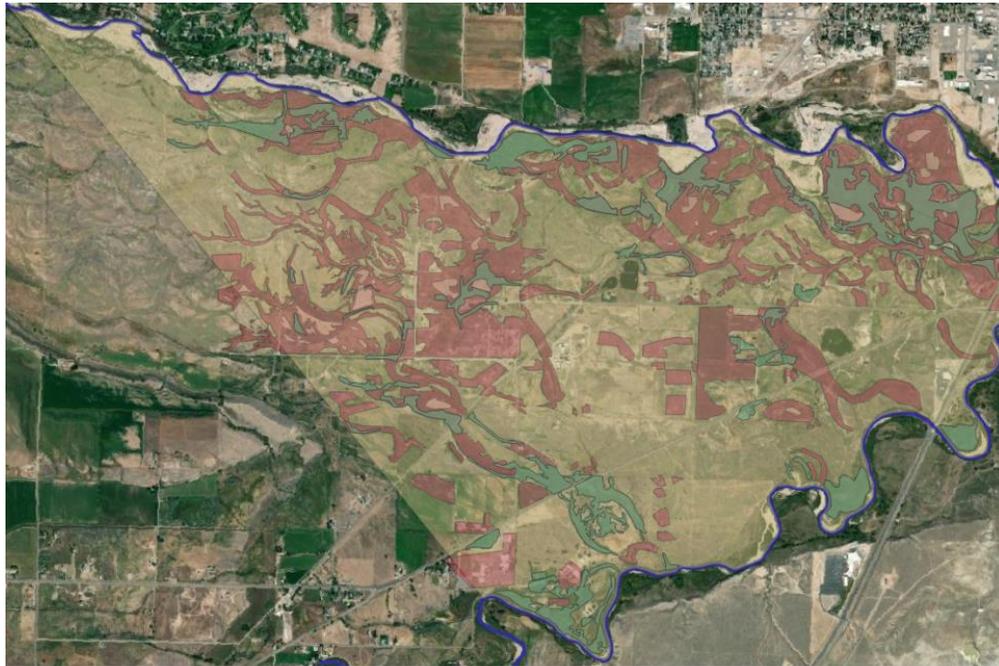
background sand (and gravel) with low LAI = 2-3
sandy with moderate cover, LAI = 4
silt and clay with high cover, LAI = 7-8

Figure 4: Map of Riverton boundaries and domain with highlighted areas assigned specified leaf area indices to deterministically assign a non-uniform recharge condition calculated as a fraction of precipitation.

3.2 Environmental Tracer Sampling

Environmental tracer concentration values were sampled during six field events for CFC12, SF6, and tritium from May 2019 through October 2020 by the UMT Hydrogeology Lab (Table 2). These tracers were sampled primarily using a peristaltic pump with Viton tubing at DOE monitoring wells. CFC12 was sampled in 250 mL glass bottles with foil lined caps. SF_6 was collected in 1L amber bottles with coned caps to avoid head space. Tritium was collected in 500 mL polyethylene bottles and capped with adequate head space. CFCs, SF_6, and stable isotopes were analyzed using a gas chromatograph at the University of Utah. Tritium concentrations were measured using a quadruple mass-spectrometer at the University of Utah (Gardner et al., Field Sampling Plan, 2020). Measured environmental tracer concentration values were used as the testing dataset when compared against simulations of MT3DMS transport.

Table 2: Field sampling dates with the number and type of environmental sample taken in Riverton, WY.

	water levels	CFC12	SF6	tritium
May 2019	5	5	4	5
July 2019	7	0	0	3
August 2019	12	10	9	6
June 2020	13	4	4	11
August 2020	9	3	3	7
October 2020	6	0	0	6

3.3 Stratigraphy and Facies Modeling

The depth to clay is different based on the interpretation of continuous versus discontinuous clay data. These interpretations are derived from uncertainty in well log stratigraphy seen in the U.S. Department of Energy's Legacy Management Geospatial Environmental Mapping System site data (DOE GEMS). Uncertainty is most apparent in well nests that have incomplete or conflicting information; specifically, where there are undifferentiated mixtures; where one well in a nest may identify clay at a specific depth, and another well shows no clay; or perhaps some wells are too shallow to reach clay but are logged in Wind River Sandstone. Therefore, the data used in kriging in the continuous-clay conceptual model (cCM) versus discontinuous-clay conceptual model (dCM) are derived from the same well log data but are differentiated by the interpreted continuity of clay and whether it exists at certain locations. This uncertainty is increased as we must interpolate between point locations to describe site stratigraphy in three dimensions. To deal with uncertainty from this spatial variability, a stochastic approach is followed (Enemark et. al., 2019) to estimate "depth to clay" and "depth to sandstone" as random variables (thereby estimating clay thickness) using the geostatistics methods outlined in the Theory section:

Distributions of empirical directional variogram parameters were built while sampling reasonable search parameters- e.g. lag distance, angle tolerance, nugget, and sill values that were observed to be reasonable during manual exploratory data analysis (EDA)- and through the use of rose diagrams (e.g. azimuth, ratio, and range). Searches for all four datasets used minimum and maximum lag distances of 150 and 430 meters, which is a little less than one-third of the maximum distance across the region. A search angle tolerance ranging from 15 and 40 degrees was chosen as this was expected to gather adequate spatial information as the search moves around the field.

As first observed during manual EDA, this process assumed geometric anisotropy, whereby different correlation lengths (variogram ranges) were collected at various search azimuths while the sill remained roughly constant. Zonal anisotropy was also observed, whereby multiple variogram sill structures were identified at different azimuths. Both of these anisotropies were

assumed to be due to the geologic nature of the data and could be accounted for in variogram modeling by randomly sampling the estimated directional variogram parameters' distributions described above, implementing these parameters in nested zonal structures to create multiple realizations of each layer's depth correlation and anisotropy. For each set of variogram parameters, kriging then generates estimated layer depths and those are saved as raster maps. The clay layer thickness is the difference between 'depth to sandstone' and 'depth to clay'.

The geometry of paleo-channels was modeled as sinusoid geobody training images as implemented by the tiGenerator utility in SGeMS. Parameters included the measured amplitude, wavelength, width, thickness, and azimuth of three channels observed in aerial photography (Boucher, 2009; Deutsch and Wang, 1996). The tiGenerator simulation populated grid cells with either an indicator for background alluvial sediment or indicators associated with one of the three observed channels- i.e., there are four possible indicators. The SNESIM algorithm then utilized spatial point locations to condition the facies' conditional distribution information to sequentially assign facies indicators at each cell of a new grid with an accurate spatial realization.

The grid used in SGeMS has the same dimension as grids used later in groundwater simulations. The grid size is ten layers thick, with 140 rows and 160 columns. The cell dimensions are 1.11m, 42.9m, 31.1m respectively. SGeMS' capability of multiple stochastic realizations was used to deal with low order geologic uncertainty in paleo channel spatial variability (Dowd, 2018), thereby providing a stochastic hydrogeologic grid framework over which to simulate fluid potential and solute transport (Tóth, 1962), with the goal of uncovering unknown subsurface structures (such as additional subsurface channels) and to assess whether subsurface paleo channel deposits acted as a conduit for groundwater flow and transport at Riverton, and test whether environmental tracer concentration and water level simulation data could distinguish these models from the continuous-clay and discontinuous-clay models described above. These methods produced mor than 221 stratigraphy- and facies-based simulation grids that were used in groundwater flow and transport simulation.

3.4 Groundwater Flow and Transport Simulation

Each alternative conceptual model is translated into numerical groundwater flow by populating a MODFLOW 2005 grid with geologic conditions, hydraulic parameters, and boundary conditions found in Table 3. The structured grid includes a DEM to represent the top boundary, kriged depth data at each layer bottom, with river locations as the model's northern and southern extent. The eastern and western boundaries are based on Thiros et al., 2021. The physical structure of the

continuous/discontinuous-clay aquifer models is assumed to be a three-layer aquifer system: alluvial aquifer, clay layer, and Wind River Sandstone. The paleo channels models have ten layers of alluvial aquifer material and use the averaged depth to clay (4.506 meters) as the bottom boundary, with no underlying Wind River Sandstone.

Hydraulic boundary conditions are assigned at the north and south boundaries in the first layer at various stress period times via the River Package in MODFLOW 2005. The boundaries at each river cell implement a constant (Dirichlet) river stage elevation estimated from a rating curve and USGS discharge data added to an incision from DEM surface elevation. The same Dirichlet river stage elevation is assigned to each layer of the alluvium cells at boundary river locations in the paleo channel conceptual models (pcCM) since each river is assumed to act as a groundwater divide. In both continuous and discontinuous-clay models, a no-lateral-flow boundary is assumed at river locations in the semi-confining clay layer. The tenth layer in the alluvial paleo channel models is the deepest alluvium and overlies the clay layer, which is assumed to be impermeable when implemented in the paleo channel structured models. The hydraulic boundary condition assigned in the sandstone aquifer is conceptualized to receive recharge from where the Wind River Formation outcrops at the surface (Love and Christiansen, 1985). This parameterization utilizes extrapolated head levels from a plane fit through field-measured water elevations measured in the sandstone, thereby assigning a constant gradient (Neumann) boundary condition to the sandstone layer.

The Recharge Package assigns a constant stress period flux across the top boundary of the domain. The uniform flux is conditioned to 25% of precipitation, historically measured at Riverton airport. Conceptual models with non-uniform recharge have a top boundary assigned with indicators to receive a percentage of precipitation (Table 2) conditioned to the plant greenness proxies. For numerical modeling of soil processes related to this study, see Thiros et al., 2021 for use of Richard's equation to solve vadose zone hydrology and its connection to groundwater flow in the alluvial aquifer at this site. In this thesis, the upper-most layer of the paleo-channeled alluvial aquifer is conceptualized as floodplain material and gets parameterized with a hydraulic conductivity value of 5 m/day, which is less than the average of the surficial aquifer (22.9 m/day). Otherwise, the hydraulic conductivity field ($\bar{K}$) of the surficial aquifer is heterogeneous, as assigned by ordinary kriging values from historic aquifer test data (DOE, 1987a), with values ranging from 5.5 to 60.1 m/day. Hydraulic conductivity of the sandstone aquifer is assumed to be homogenous and is assigned the mean estimated value from aquifer tests in the sandstone aquifer,

1.909 m/day. The clay layer is assigned a constant hydraulic conductivity of 0.0001 m/day. Cells indicated as paleo channels are assigned a higher hydraulic conductivity than the kriged surficial aquifer test data. See Table 3 for additional parameter information including the storage coefficient values assigned to each layer. All layers are simulated as unconfined aquifers.

MT3DMS transport simulations are used to solve environmental tracer transport and fate through the prescribed MODFLOW models. The top transport boundary condition uses historically converted atmospheric to aqueous concentrations (Thiros et al., 2021) to enter the groundwater system with recharge (Figure 5). Solute boundary conditions at river reach cells in the three-layer models assume an aqueous concentration in constant equilibrium with the atmosphere (Thiros and Gardner, vocal communication). Initial conditions for environmental tracer concentrations in clay and sandstone are zero, justified by rivers acting as groundwater divides- in addition to the extent of the clay beyond the rivers being unknown. Diffusion and dispersion parameters were estimated using activation energy relationships developed by Zheng et al., 1998 for CFC12; and King and Saltzman, 1995 for SF6. These coefficients are the same for all layers. Processes including sorption or biodegradation are not simulated.

For each conceptual model, there are 221 numerical groundwater flow and transport simulations corresponding with the stochastic grids to estimate water levels and environmental tracer concentrations. The simulations start in 1942 and finish in December 2020, with an initial hydraulic condition starting 0.2 meters above land surface. The first 230 stress periods, which finish in December 2018, vary in length by a prescribed number of days that corresponds with the first four months of a year, followed by the next three months, followed by the last five months of the year, i.e., 120 days, 92 days, 153 days. The remaining 24 stress periods start in January 2019 and run to December 2020 with each month the length of days that corresponds to that month, e.g. 31 days in January, 28 days in February, etc.

Table 3: Conceptualization and model summary table for physical and process hydrogeologic structures, and parameters specific to groundwater flow and transport process simulations. Parameters include site-specific geologic conditions, hydraulic parameters, field-measured and estimated flow and transport boundary conditions.

Conceptual Flow: physical structure				
layer	**Gravel and Sand**	**Clay**	**Sandstone**	**Paleo Channels**
dimensions	3 x 140 x 160			10 x 140 x 160
mass balance	transient	transient	transient	transient
hydraulic conductivity /complexity	heterogeneous[1] /isotropic	homogeneous[2] /anisotropic	homogeneous[3] /isotropic	heterogeneous[4] /isotropic
storage	0.205	0.190	0.200	0.205
river boundary conditions	Dirichlet/estimate river stage	no flow	Neumann/water levels	Dirichlet/estimate river stage
initial conditions	0.2m above surface	surface elevation	surface elevation	0.2m above surface

Conceptual Flow: recharge process structure			
uniform recharge	**spring/summer**	**winter**	
	25% Precipitation (P)	2.5% P	
non-uniform recharge	**low LAI**	**mid LAI**	**high LAI**
	29.85% P	23.33% P	13.89 % P

Transport Model			
	Gravel and Sand/PCs	**Clay**	**Sandstone**
tracer source	recharge and rivers	premodern	premodern
porosity	0.33	0.25	0.3
long. disp.[5]	14.1	1	7
ratio of horiz. trans. disp. to long. disp.	0.15	0.075	0.125
ratio of vert trans disp. to long. disp.	0.035	0.035	0.035

Tracer - Specific Parameters			
	CFC12	**SF6**	**tritium**
diffusivity coefficient[6]	6.76E-057	7.618E-05	-

---

[1] Kriged hydraulic conductivity from field aquifer tests with a mean of 22.995 meters per day.
[2] Constant and uniform parameter with mean value of 0.0001 meters per day.
[3] Constant and uniform parameter with mean value of 1.9095 meters per day.
[4] Kriged hydraulic conductivity with indicated paleo channel cells randomly assigned hydraulic conductivity of 30, 50, or 69 meters per day.
[5] Longitudinal Dispersivity from https://www3.epa.gov/ceampubl/learn2model/part-two/onsite/longdisp.html
[6] From the relationships developed by Zheng et al., 1998 for CFC12 and King and Saltzman, 1995 for SF6, the median value is calculated using a temperature range: 7 to 23 degrees Celsius
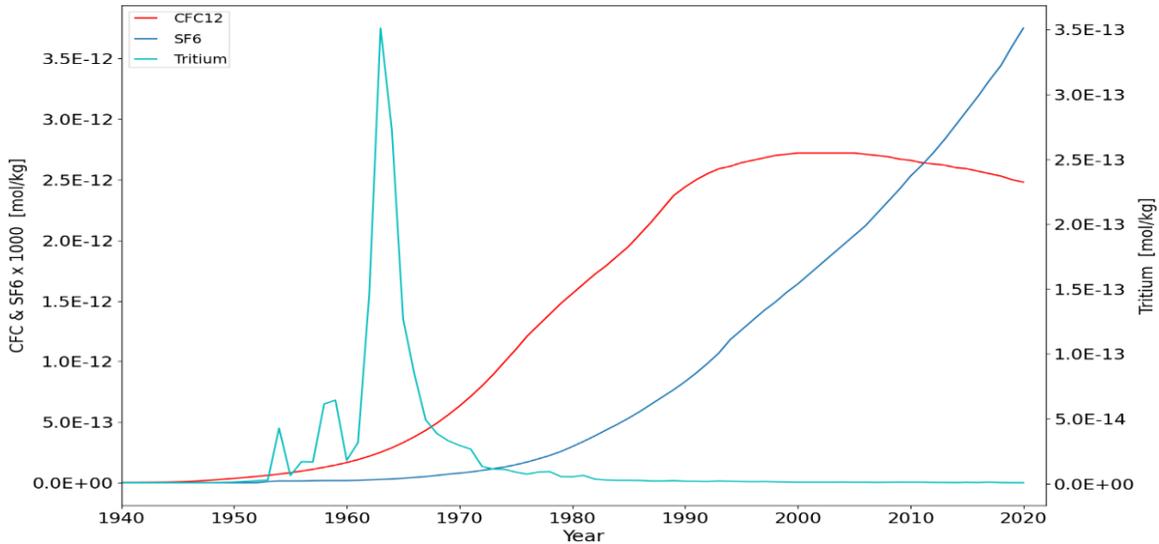
Figure 5: Aqueous environmental tracer concentrations condition applied to groundwater recharge and river locations.

3.5 Post Processing

Simulated water level results at an elevation of zero or that were negative were removed from consideration. This filter also applies to simulated tracer concentrations. Therefore, there are 35 model output samples with joint locations and times (sample points) with reasonable water levels and concentrations that align with field sampling events that took place between May 2019 and October 2020 (Table 2). Simulation output from each sample point is stored in arrays representative of each of the six conceptual models (CMs). Qualitative and quantitative analysis between models will aid in distinguishing between and observing performance of the conceptual models; after all, conceptual models are both a qualitative and quantitative description (Neuman and Wierenga, 2003), so testing CMs deserves both qualitative and quantitative assessment. Comparing simulations to field-measured data will attempt to validate one or more CMs. Lastly, synthetic sample locations are chosen to partition alluvial groundwater into upgradient and downgradient zones, with downgradient wells located south of the former contaminant source zone (Figure 6). This analysis will be used to make physical interpretations related to mixing between aquifers and is simply based on the direction of simulated water levels and environmental tracer concentrations' stochastic distribution from a given CM compared to the response of another CM.
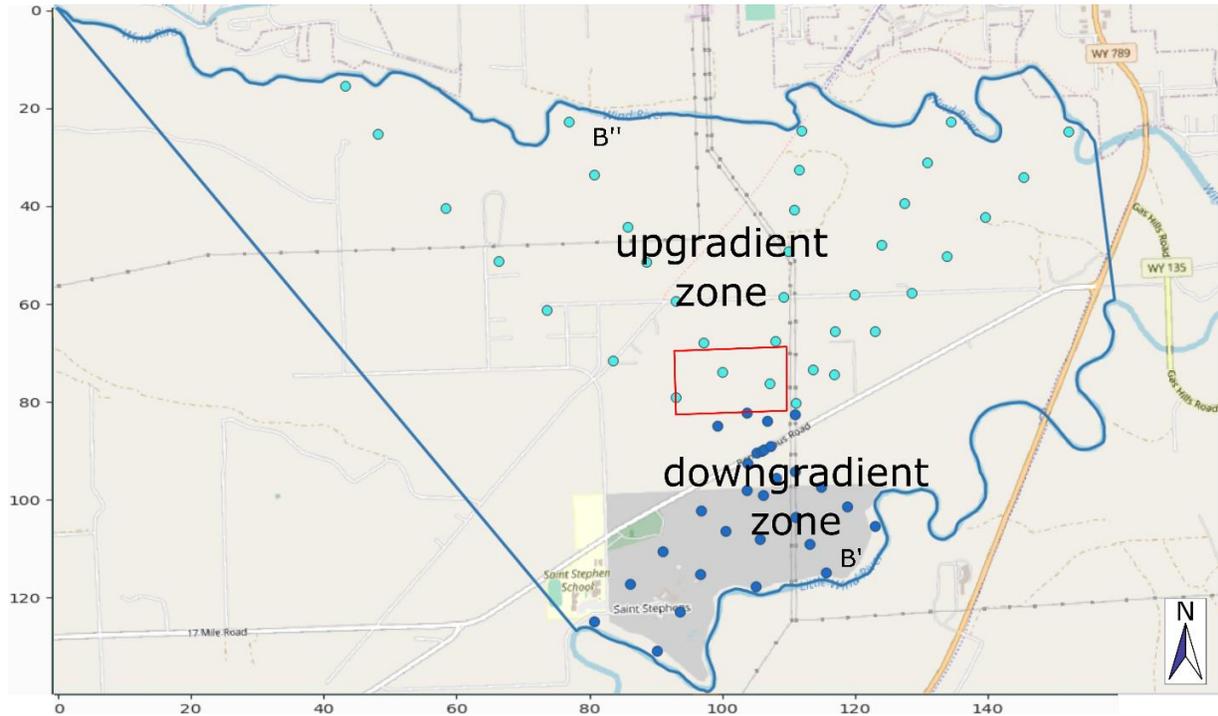
Figure 6: Numerical model sample locations along synthetic cross sections in upgradient and downgradient zones. The red box is the former contaminant zone. Nine sample locations on two different roads (Rendezvous Rd and Goes in Lodge Rd) were removed from consideration. There are twenty synthetic locations in the downgradient zone and thirty-six in the upgradient zone.

The first qualitative assessment is an inter-model visual comparison for simulated water levels and environmental tracer concentrations that have been normalized at individual sample points along cross section A to A' (Figure 2). In this method, differences in the shape of lumped kernel density estimates of water levels and simulated tracer concentrations are visually compared. Due to non-normality observed in these distributions, both at individual locations and later when all 35 sample points are lumped together, the Wilcoxon Rank Sum test (Bauer, 1972) is used to numerically assess differences in each of the conceptual models' simulation distributions. The test's null hypothesis is that two sets of sample measurements are drawn from the same distribution, with the alternative that one distribution may be shifted to the left or right of the other (two-sided test). Similarly, a two-sample Kolmogorov-Smirnov test is applied to each CM comparison and tests a null hypothesis that the cumulative distribution function (CDF) for one CM distribution is the same as the CDF from an alternate CM. The alternative hypothesis is that they are different. Levene's test for variance homogeneity is used to assess equivalence of inter-model simulation distributions in terms of their variance. The last method used to compare CMs computes Mahalanobis distances (MHDs) between simulation results to aid in evaluating the ability of simulated well water levels and tracer concentrations to distinguish between models. An assumption is that larger MHDs (between CMs) are related to an environmental variable's

21

ability to distinguish better than another. The smallest MHD among four variables means that variable has less ability to distinguish between CMs and, therefore, to reduce CM uncertainty.

All of the above calculations will utilize data that have undergone a linear rescaling procedure, whereby, each respective variable from individual sample points is scaled between zero and one:

$$x' = \frac{x - \min(\vec{x})}{\max(\vec{x}) - \min(\vec{x})}$$

where x is the original vector of simulation data from the sample point for a given CM, $\vec{x}$ is the set of all six CMs' results at that sample point, and x' is the normalized vector. Then, the normalized results from all 35 sample points are combined as one standardized distribution assigned to each CM for the analysis.

Lastly, MHDs between non-normalized simulation data and field-measured data are calculated at each sample point in an attempt to validate one of the six CMs. This validation assumes that smaller distances indicate a better match to field data, while larger distances indicate a worse match. Due to order-of magnitude differences associated with each variable's unit of measure, the resulting MHDs must again be linearly transformed. This will take summed MHDs, respectively per variable and per CM, and normalize to the minimum and maximum MHD from all CMs. Levene's test for variance homogeneity is used to assess intra-model variance among the four simulated variable's distances to field data as a way to assess each variable's sensitivity to field data.

4. Results

4.1 Geostatistical Simulations

The mean thickness from 450 realizations for continuous clay is 0.600m and the mean clay thickness from 450 realizations for discontinuous clay is 0.530m. Each of the continuous and discontinuous-clay models respectively have a similar range of thicknesses between 0 and 2.25 meters. Figure 7 shows two examples of simulated clay layer thickness maps, one for continuous and one for the discontinuous clay conceptualization. Also shown in the figure are two examples of coherent channel structures in the second layer of the resulting SNESIM realizations that are correctly mapped to observed paleo channel locations. Discrete and sinuous channels that were modeled as TIs are not present in the multiple-point statistics realizations. Rather, SNESIM realizations are inundated by geobody indicators whose features have a wide and somewhat random structure. Although, this is conceptually representative of fluvial channel deposits with

non-discrete boundaries. These indicators for channel deposits are found exclusively in the second layer of simulation grids, which is also representative of paleo-channels acting as principal conduits to groundwater flow in the alluvial aquifer. The other SNESIM grid layers do not have these distinct channel representations. Rather, the third layer has smaller string-like channel indicators under the principal shallow channels, which can be representative of the channel bottom, and blocky features with no coherent structure occupy the remaining seven layers. The grids produced using this method were directly used in groundwater flow and transport models, with elevated hydraulic conductivity values assigned to areas with facies indicators.
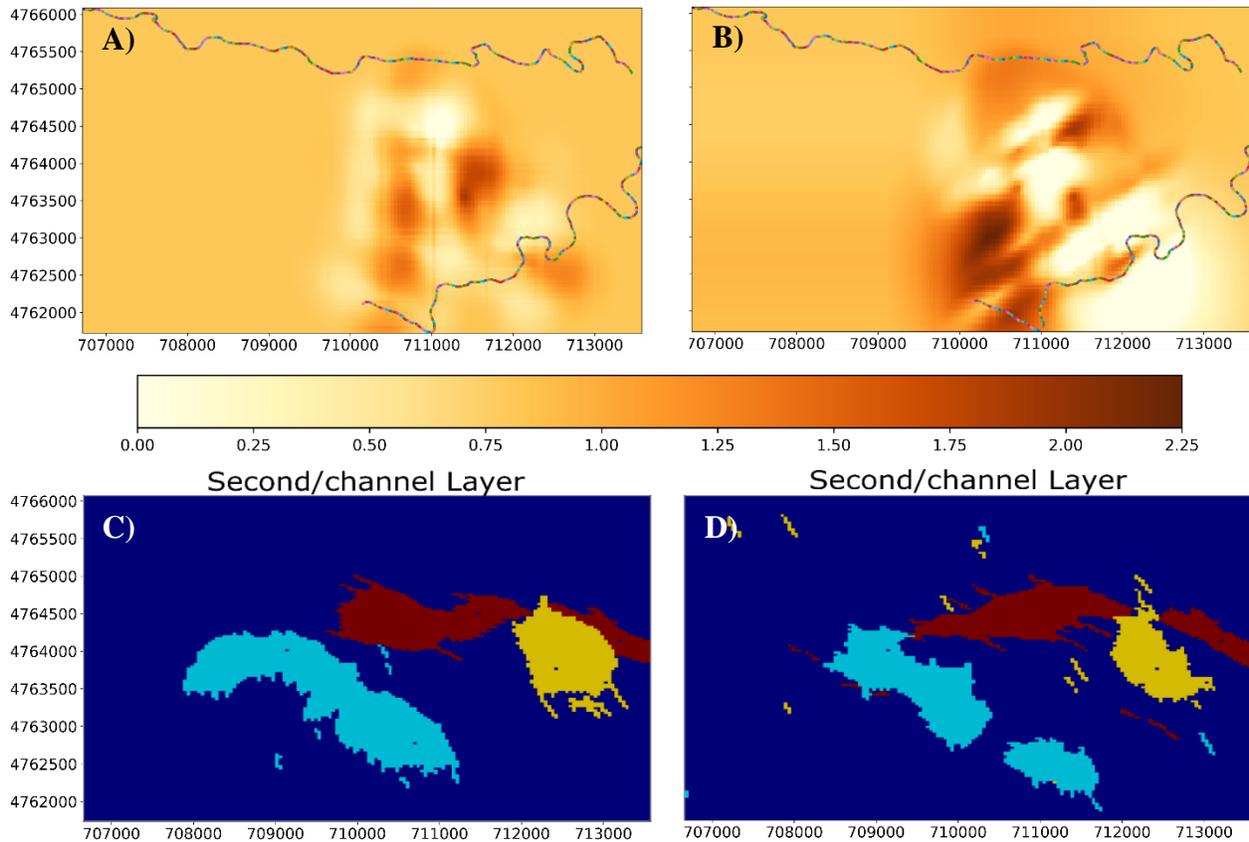
Figure 7: A) Map view of one stochastic realization of continuous-clay thickness (in meters). B) Map view of one stochastic realization of discontinuous-clay thickness (in meters). The ribbons in A) and B) are boundary locations for the northern Wind River and southern Little Wind River. C) and D) Are two examples of stochastic SNESIM grid layer realizations with three indicators of different colors that represent possible paleo-channel configurations in the alluvial aquifer. Only the second layer from the SNESIM grids are shown as they were the only layers that had reasonable channel-like structures.

4.2 Inter-model Comparison

Gaussian kernel density estimates for comparing normalized simulated water levels lumped at all 35 sample points for all six CMs are shown on the y-axis of Figure 8 with environmental tracer concentrations CFC12, SF6, and tritium on a new x-axis. Means and standard deviations for these data are summarized in Table 4 A). However, due to the non-normality of the data, p-values derived from Levene's tests use the absolute deviation from median values and are summarized

in Table 4 B). Kernel density estimates for normalized simulated water levels and environmental tracer concentrations at various sample points associated with Riverton monitoring wells along cross section A to A' are shown in Figure 9. In both figures 8 and 9, there are distinct 'puddles' of simulated variables in normalized space associated with each of the six CMs. For example, at this visual level, there are depictable differences in distribution shape (peaks, tails, and some outliers) of simulation data between continuous-clay, discontinuous-clay, and paleo channel simulations. There appears to be little difference in simulated water level response when comparing any pair of CMs that are only different by their alternate recharge process structure. There are also similarities in the general shape of simulated environmental tracer concentration distributions between pairs of CMs that differ by their alternate recharge structure, although one distribution might be shifted from the other (see Figure 9). Water levels, CFC12, and SF6 concentrations simulated in paleo channel models (pcCM and pcnCM) often exhibit a wider and unanimously lower distributions than in the three-layer systems.
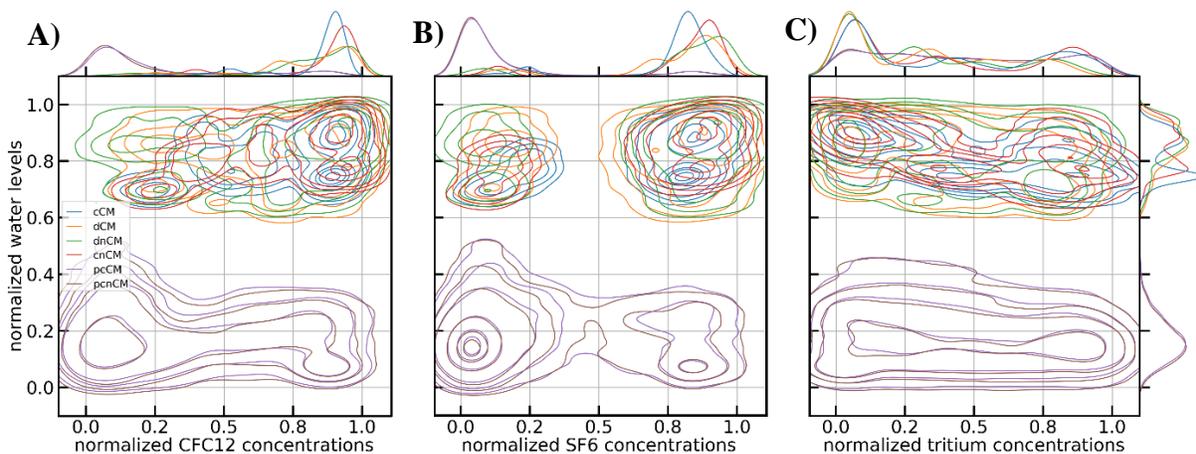
Figure 8: Kernel density estimates for lumped normalized simulation results. Normalized water levels on the Y-axis, and A) shows CFC12, B) SF6, and C) tritium concentrations on each respective X- axes.

To navigate the Levene's analyses in Table 4 B) and Kolmogorov-Smirnov in Table 5), for water levels and SF6 results, start with the row CMs and compare against CMs that are in the columns with results being above the diagonal. Results are transposed for CFC12 and tritium: for these, start with a CM in the column and compare against a CM in a row with results being below the diagonal. All comparisons start with cCM vs. dCM and end with pcCM vs. pcnCM.

Several observations worth mentioning for later physical interpretations is that water levels in discontinuous-clay models (dCM and dnCM) are bimodal at local sample points (Figure 9), and have a standard deviation of 0.077 and 0.079 at the lumped field scale, which are wider than water

24

level distributions in the continuous-clay models (cCM and cnCM), which have standard deviations of 0.076 and 0.076, respectively, in normalized space (Table 4 A)).
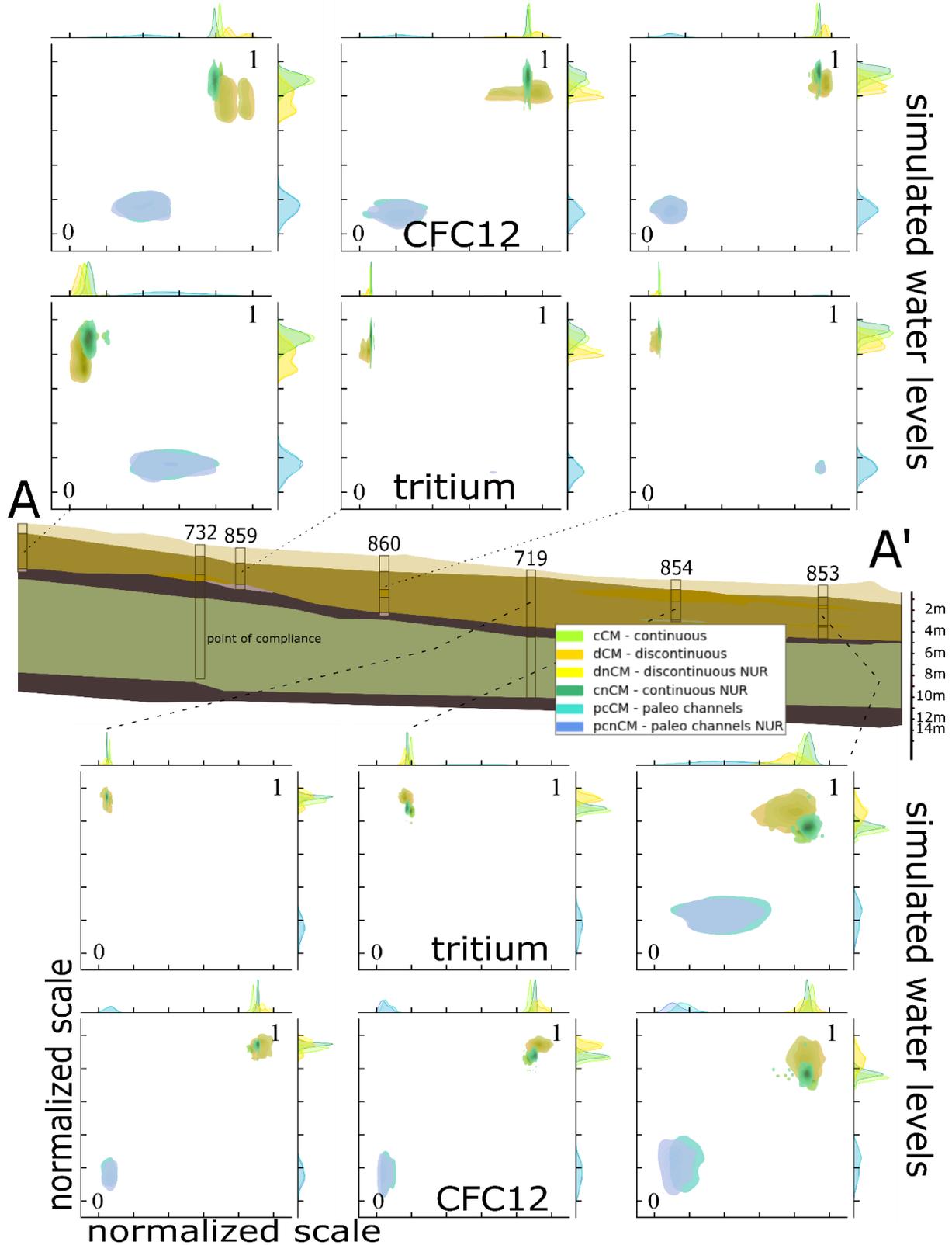
Figure 9: Cross section along A to A' with normalized simulated water levels on vertical axes and normalized tracer concentrations on the x-axes for tritium and CFC12.

Levene's test implies that these water level distributions in the discontinuous-clay models are significantly wider than those in the continuous-clay models, with the greatest p-value among the four comparisons of 2.3 x 10^{-7} for dnCM vs cnCM. Further, as seen in Levene's tests, the absolute median residual (~variance) for any variable simulated in the cCM is not the same as any other model (Table 4 B)). By the same reasoning (larger standard deviations and small p-values to indicate there are differences in variance), the ranges of simulated CFC12 concentrations in models with discontinuous clay are larger than those of CFC12 concentrations from continuous-clay models. Conversely, for tritium, the range of simulated tritium concentrations in the dCM and dnCM is smaller than that of tritium concentrations from models with the continuous clay layer. The distributions themselves of tritium concentrations in the discontinuous-clay models are also lower than (shifted left from) tritium distributions from the continuous clay models.

Table 4: A) Calculated mean and standard deviation of the normalized modeled variables associated with the 35 field sampling times. B) P-value results from Levene's test for variance homogeneity for inter-model CM comparisons. There are two sets of comparisons in each table (so that all pairs of CMs are compared for each of the 4 variables) either in the lower or upper triangular sections; all comparisons start with cCM vs. dCM and end with pcCM vs. pcnCM. The minimum p-value is < 2.2 x 10^{-16}.

A)	water levels		CFC12		SF6		tritium	
	mean	SD	mean	SD	mean	SD	mean	SD
cCM	0.818	0.076	0.816	0.173	0.757	0.204	0.410	0.333
dCM	0.840	0.077	0.780	0.204	0.762	0.234	0.320	0.272
dnCM	0.859	0.079	0.792	0.227	0.792	0.264	0.313	0.282
cnCM	0.844	0.076	0.828	0.200	0.786	0.242	0.395	0.334
pcCM	0.175	0.080	0.259	0.273	0.156	0.236	0.454	0.308
pcnCM	0.167	0.080	0.254	0.281	0.152	0.242	0.453	0.311

B) — Water Levels (upper triangle) / CFC12 (lower triangle)

CFC12 ↓	cCM	dCM	dnCM	cnCM	pcCM	pcnCM
cCM	—	2.2 x 10^{-16}	2.2 x 10^{-16}	1.9 x 10^{-9}	2.2 x 10^{-16}	2.2 x 10^{-16}
dCM	2.2 x 10^{-16}	—	0.304	1.7 x 10^{-10}	0.177	0.459
dnCM	2.2 x 10^{-16}	0.969	—	2.3 x 10^{-7}	0.655	0.684
cnCM	9.0 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	—	4.1 x 10^{-6}	9.7 x 10^{-8}
pcCM	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	—	0.548
pcnCM	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	0.301	—

B) — SF6 (upper triangle) / tritium (lower triangle)

tritium ↓	cCM	dCM	dnCM	cnCM	pcCM	pcnCM
cCM	—	2.2 x 10^{-16}	2.2 x 10^{-16}	6.9 x 10^{-12}	8.0 x 10^{-14}	5.3 x 10^{-12}
dCM	2.2 x 10^{-16}	—	0.232	9.6 x 10^{-9}	1.1 x 10^{-7}	1.8 x 10^{-8}
dnCM	2.2 x 10^{-16}	0.426	—	7.9 x 10^{-11}	9.7 x 10^{-10}	1.5 x 10^{-10}
cnCM	0.003	2.2 x 10^{-16}	2.2 x 10^{-16}	—	0.632	0.928
pcCM	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	—	0.699
pcnCM	2.2 x 10^{-16}	2.2 x 10^{-16}	2.2 x 10^{-16}	1.6 x 10^{-15}	0.204	—

Using four of the Levene's tests, there is no evidence against variance homogeneity among all simulation variables (p-value = 0.304 [water levels], 0.969 [CFC12], 0.232 [SF6], and 0.426 [tritium]) between the two discontinuous-clay models. Similarly, there is no evidence that any variance of the four variable's distributions are significantly different between the two paleo-channel models (p-value = 0.548, 0.301, 0.699, 0.204). The variances of water level distributions from either of the dCM and dnCM are not significantly different from distribution variances from either paleo-channel model (p-value = 0.177 [cCM vs. pcCM], 0.459 [dCM vs. pcnCM], 0.655

[dnCM vs. pcCM], 0.684 [dnCM vs. pcnCM]). There is no evidence that the variance of the normalized SF6 concentrations in the cnCM is significantly different from the variances from either the pcCM or pcnCM (p-value = 0.632, 0.928). However, as discussed below, based on results from the Kolmogorov-Smirnov tests (Table 5), none of the CDFs of these functions just described are the same. The remaining tests in Table 4 B) (there are 46) all provide evidence for the alternative to indicate variance heterogeneity between each alternate simulation distribution. These 46 tests primarily compare CMs with alternate physical structures, i.e., continuous vs. discontinuous vs. paleo channels (or variations of this with non-uniform recharge), and/or are the comparison between the two continuous-clay models.

Based on the ability of this test to determine when there are differences between variances of different distributions, e.g., the p-value for cCM vs. cnCM for water levels is 1.9×10^{-9}, while both of their 2-decimal place standard deviations are 0.076; and based on visually wider tracer concentration distributions compared to water levels in Figure 8, these results can generally show that simulated tracer concentrations span a wider range of estimates in normalized space compared to water levels.

Preliminary Wilcoxon Rank Sum p-value analyses for lumped simulation data showed an overwhelming ability for each variable to distinguish between models so are not presented here. Results from a subsequent two-sided two-sample Kolmogorov-Smirnov analysis (Table 5) also show there are overwhelming differences between all the CMs, except for the comparison of simulated tritium concentrations between the pcCM and pcnCM, where there is no evidence of a difference between the distributions.

Table 5: P-value results from Kolmogorov-Smirnov tests for inter-model CM comparisons. The null hypothesis is that there is no difference from one model's CDF to the one it is compared against. The alternate hypothesis is that the CDFs are different. These results are from the 35 lumped sample point distributions and are calculated respectively for each environmental variable.

| | | Water Levels | | | | | | SF6 | | | | | |
		cCM	dCM	dnCM	cnCM	pcCM	pcnCM	cCM	dCM	dnCM	cnCM	pcCM	pcnCM
CFC12	cCM	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
	dCM	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
	dnCM	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}
	cnCM	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	2.2×10^{-16}	2.2×10^{-16}
	pcCM	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	1.0×10^{-5}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	—	5.8×10^{-9}
	pcnCM	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	1.2×10^{-10}	—	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	2.2×10^{-16}	0.803	—

The SF6 columns in rows for tritium are labeled with the row axis **tritium**.

Similar to the above comparisons, the Mahalanobis distance (MHD) is calculated between each of the distributions resulting from the 35 lumped sample points (Figure 10). The distances are

ranked from smallest to largest with a log scale on the y-axis. Models with identical physical geologic structure but that are different by the applied alternate recharge have the smallest MHDs between them. These three comparisons are on the left side of the plot. In terms of CM uncertainty, these are associated with uncertainty related to alternate conceptualizations of the process structure of these models. The next four (middle-left) comparisons are associated with continuous- versus discontinuous-clay uncertainty of the physical structure. On the right side of the plot, these largest MHDs are calculated between either continuous- or discontinuous-clay models versus paleo channel models. Accordingly, these comparisons are associated with CM uncertainty of whether the subsurface has a continuous and/or discontinuous clay layer between the unconsolidated alluvium and consolidated sandstone aquifers, versus whether the alluvial aquifer has paleo channels through the alluvial aquifer. As stated in the Methods section, in terms of conceptual model uncertainty, larger MHDs are assumed to be related to an environmental variable's enhanced ability to distinguish between CMs, and therefore that variable has a better ability to reduce conceptual model uncertainty.

The far-left side of the plot has water levels with the closest distance, and all three tracers measure greater distances between these models. The middle-left of the plot shows SF6 and water levels as the closest simulation variables (smallest MHD) among all four of these models, with tracers CFC12 and tritium measured the greatest difference. On the right side of the plot, SF6 shows the furthest distance between models, while normalized simulated tritium concentrations are closest between these models.
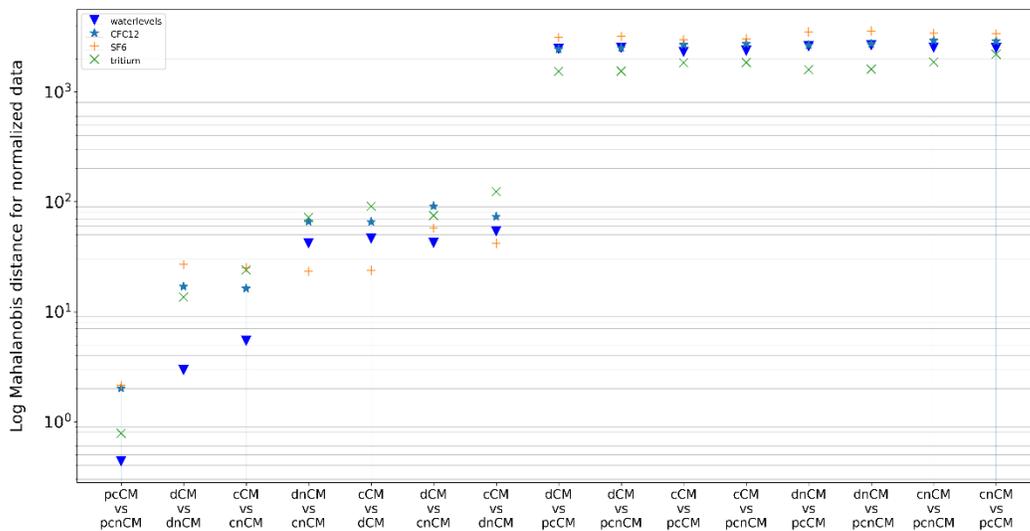
Figure 10: Log-scale Mahalanobis distance between conceptual models.

28

For the synthetic upgradient and downgradient well locations, a similar post-processing filter to the one described above was applied for removing simulation locations with unreasonable results. Results for paleo channel models at upgradient locations not shown due to one or more tracers' concentrations being multiple orders of magnitude these models compared to the other models. From the remaining locations, for which there are 24 lumped upgradient (Figure 11 A)) and 18 lumped downgradient locations (Figure 11 B)), kernel density estimates of normalized CM results from August, 2019 can again be used to infer that unique characteristics seen in distributions can only be due to differences that originated during conceptual model development, and those differences propagated through to the resulting numerical groundwater flow and transport models. The kernel density estimates show upgradient locations in the cCM and cnCMs to have simulated a water table higher than in models with the discontinuous clay structure. At downgradient locations, water levels are higher in the dCM and dnCM compared to any other model. Aside from outliers, all three tracers appear to have more left skewed results at downgradient locations. This leftward skew is more apparent in the models with a discontinuous clay structure compared to those with continuous clay (Figure 11 B)).
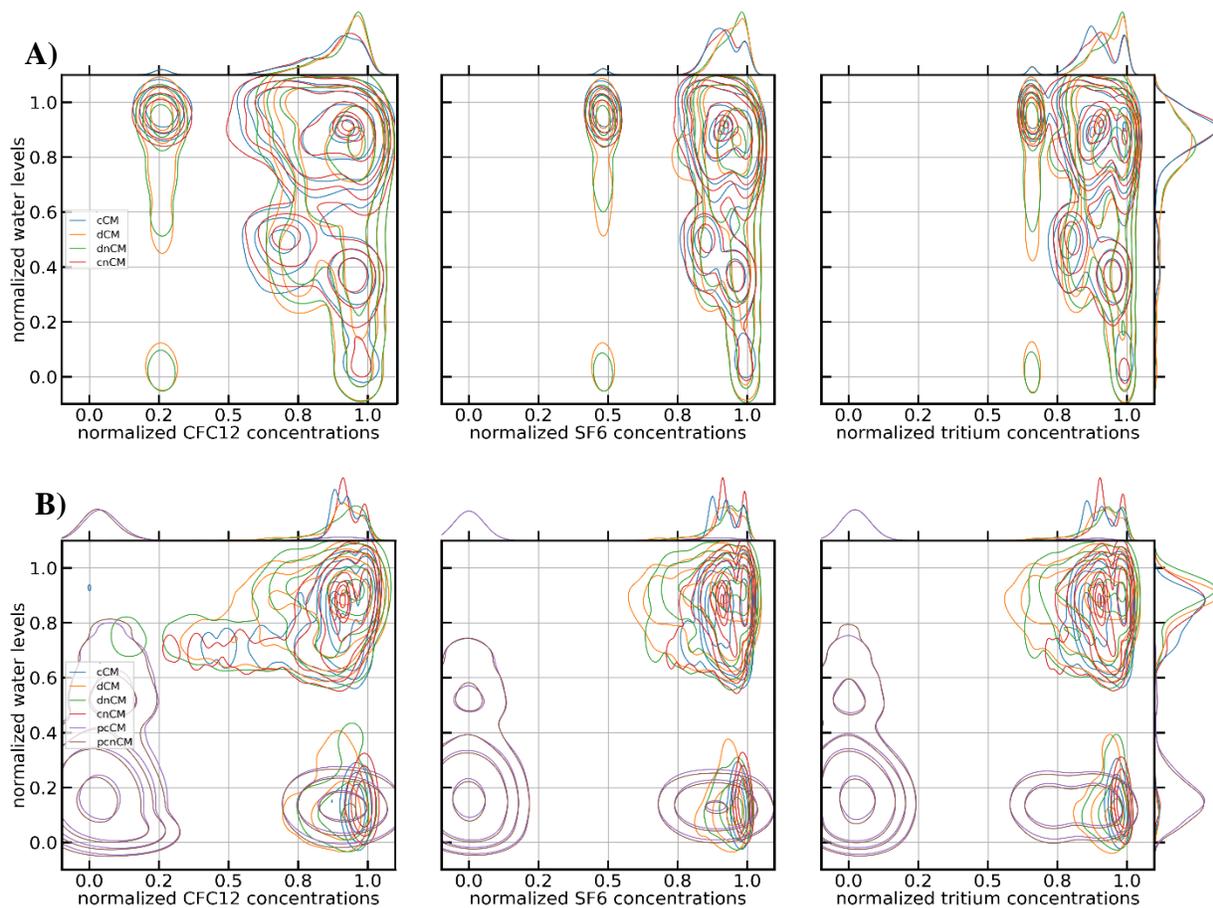
Figure 11: Kernel density estimates for lumped normalized simulation results from A) upgradient and B) downgradient locations.

4.3 Model Validation

The distance between each simulated variable and field measurement is quantified with the Mahalanobis distance to assess the ability of uncalibrated CMs to reproduce field measurements. Because there was no model calibration, no model is expected to match field-measured data. All four of the three-layer CMs are a similar distance away from field-measured data ([Figure 12]). Tritium concentrations in those four models are closest to field-measured observations. Simulated CFC12 concentrations in those four models have the second smallest Mahalanobis distance from field-measured concentrations. Simulated SF6 concentrations are third closest to field-measured concentrations, and water levels are furthest from the field-measured water table.
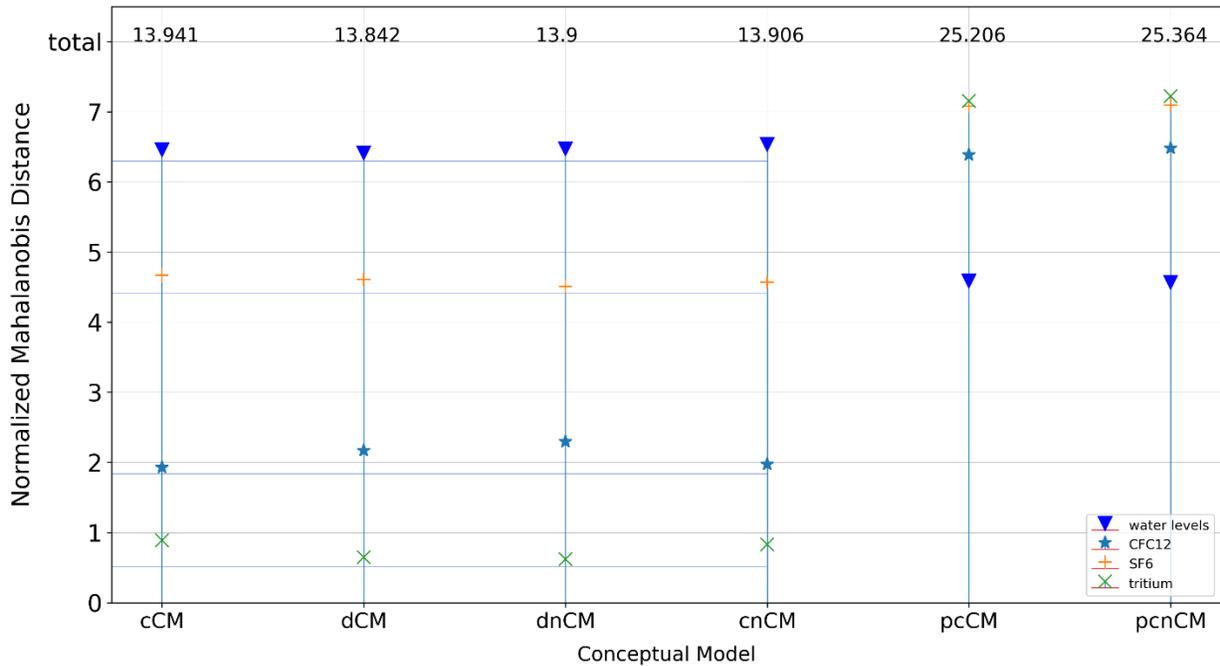
Figure 12: Sum of the Mahalanobis distances between simulation data and field samples.

Table 6: P-value results from Levene's test for variance homogeneity for Mahalanobis distances between CM simulations and field-measured observations.

| | | Water Levels | | | | | | | SF6 | | | | | |
		cCM	dCM	dnCM	cnCM	pcCM	pcnCM		cCM	dCM	dnCM	cnCM	pcCM	pcnCM
CFC12	cCM	—	0.980	0.990	0.986	0.571	0.567	tritium	—	0.957	0.873	0.916	0.432	0.428
	dCM	0.556	—	0.989	0.966	0.585	0.581		0.518	—	0.918	0.961	0.410	0.407
	dnCM	0.823	0.679	—	0.977	0.576	0.572		0.240	0.549	—	0.955	0.357	0.354
	cnCM	0.696	0.774	0.866	—	0.560	0.556		0.658	0.860	0.470	—	0.377	0.373
	pcCM	2.8×10^{-4}	7.3×10^{-4}	3.2×10^{-4}	3.5×10^{-4}	—	0.995		6.0×10^{-4}	6.4×10^{-4}	6.7×10^{-4}	6.3×10^{-4}	—	0.997
	pcnCM	2.4×10^{-4}	6.4×10^{-4}	2.8×10^{-4}	3.0×10^{-4}	0.995	—		6.1×10^{-4}	6.4×10^{-4}	6.7×10^{-4}	6.30×10^{-4}	0.991	—

Among these four models, there appears to be a (visual) tendency for greater variability among the tracer concentrations with respect to field-measured data, while the water levels are hardly sensitive to the field data. However, from Levene's test, none of the MHDs' variability is significantly different with respect to any of the other variables based on a lower confidence limit p-value of 0.05, except between paleo channel models and the other models. It can be observed from Table 6 that, in general, p-values for all comparisons for tracers have lower p-values than comparisons for water levels.

The paleo channel models are, respective to each other, a similar distance away from field-measured data. There is an improvement in these CMs to simulate lower water levels more accurately; however, the environmental tracer concentrations do not perform as well in these models at predicting the field measurements. Simulation data from the discontinuous-clay CM is closest to field-measured. Paleo channels with uniform recharge are the furthest for field-measured data.

5. Discussion

The kriged clay thickness and SNESIM of paleo channels offered acceptable representations of the conceptual physical structures in the context of using conditioning data (e.g. well log description) in an attempt to model multiple conceptual physical subsurface structures, with inclusion of a stochastic approach to assess CM uncertainty only. The thickness of the modeled clay layer was 0.07 meters less in the discontinuous-clay model compared to the continuous-clay ($n = 450$). Realized paleo channel grids from SNESIM may not have modeled channels as discretely sinuous bodies, but rather as widespread random bodies; however, conceptually, the representations in Figure 7 may represent channel deposits left behind as former rivers migrated across the land.

The purpose of this research was to test and assess whether simulating environmental tracer mass transport in groundwater flow and tranport processes improves the ability to distinguish between multiple hydrogeologic CMs, and if tracers are as useful or better than using hydraulic data in model validation. The analyses of water level and environmental tracers' kernel density estimates in normalized space allowed a visual analysis of CM uncertainty: overlap of distributions indicate non-uniqueness and no reduction in uncertainty; conversely, unique peaks and tails of the simulated variables indicate differences between models that can only be explained by uniqueness in the conceptual models' structures. Where these differences exist, there is a reduction in

uncertainty. With few exceptions, the results of the Kolmogorov-Smirnov tests showed that all variables had an overwhelming ability to detect differences in simulation results at field-scale resolution. Four exceptions occurred where the minimum possible p-value ($< 2.2 \times 10^{-16}$) did not occur, was for comparisons between paleo channel models: there was no evidence of a difference in tritium distributions between pcCM and pcnCM (p-value = 0.803), the smallest p-value was for CFC12 to indicate the greatest evidence of a difference (1.2×10^{-10}), the next smallest p-value was for SF6 (5.8×10^{-9}), with there still being a difference in distributions for water levels (p-value = 1.0×10^{-5}). By the general assumption that a smaller p-value indicates greater evidence for the alternative hypothesis, environmental tracers CFC12 and SF6 performed better at distinguishing between the two paleo-channel models than water levels or tritium. Levene's test for variance homogeneity confirmed there to be no evidence of a difference in simulation distributions' variances between models that were different by the applied recharge conceptualization. Visual observation of Figure 9 agrees and shows that the shape of these distributions (model with uniform recharge vs. model with non-uniform recharge) are similar, although one distribution might be shifted in one direction with respect to another. However, the Kolmogorov-Smirnov test found that nearly all of these distributions were not the same.

The intramodel variance of normalized estimates of simulated water levels is smaller than the variance of tracer concentrations, i.e., tracers have a wider range of estimates in the normalized space, which means simulated tracers may have a poor estimation power compared to using water levels. Also in terms of intramodel variance, the Levene's test found no statistical differences in the variability of MHDs calculated between field and simulation data. There is only marginally smaller p-value estimates for the tracers' information compared to water levels. Still, statistically, there is no evidence to support any claim that any variable should be preferred over another because it is more sensitive or useful to compare to field data.

From the tools used in this study, the Mahalanobis distance (MHD) calculations provided a good quantitative metric to test which, or whether, any environmental variable has the best ability to reduce conceptual model uncertainty. It is certain from intermodel MHDs of any variable that paleo-channel models are easily distinguishable from the other models. For the variety of comparisons between cCM, dCM, dnCM, and cnCM, which are related to uncertainty of clay continuity, the largest MHDs for tritium and CFC12 means these variables had the greatest ability to distinguish and reduce uncertainty between these conceptualizations, while SF6 and water levels were less able to distinguish between these models. Then, for uncertainty related to the

alternate recharge conceptualization, water levels performed the worst at distinguishing between models while tracers had better abilities to reduce this uncertainty.

In terms of model validation, the Mahalanobis distance calculation showed tritium and CFC12 closest to field-measured results in the three-layer model with the discontinuous clay. Models conceptualized with transmissive paleo channels in the alluvial aquifer had simulated water levels closest to the field measurements, however, their totaled MHDs of all variables were furthest away from field data.

5.1 Physical Interpretations

Preliminary geologic characterization efforts showed that the Wind River Formation widely outcrops across the region outside of the chosen model domain. A hypothesis related to the purpose of this modeling study is that if groundwater recharge is greater to Wind River sandstone at outcrop than to the alluvium, that could cause artesian processes at regional or at model scale. Then, if the clay separating the sandstone aquifer from the alluvial aquifer is more discontinuous than it is continuous, upwelling could occur from sandstone into the surficial aquifer. Then, perhaps, more extreme upwelling could cause the remobilization of contaminats during higher recharge events.

The results showed water level at lower elevations in surficial aquifer groundwater at upgradient locations in models with a more discontinuous clay layer than the models with continuous clay. An explanation is that groundwater in the alluvial aquifer is able to downwell into the sandstone aquifer. Water levels simulated at downgradient locations are higher in models with the more discontinuous clay layer compared to models with continuous clay. This may test the above theory; the discontinuous clay, along with the other assigned boundary conditions and domain properties, emulated an upwelling process in this area with older waters moving into the alluvial aquifer. Results of environmental tracer mass transport may further support this theory. As previous reported, environmental tracer concentrations in downgradient locations tended to be lower in discontinuous-clay models than in continous-clay models. This can be explained by simulated flow of premodern groundwater upwelling from the Wind River sandstone and mixing with alluvial aquifer groundwater.
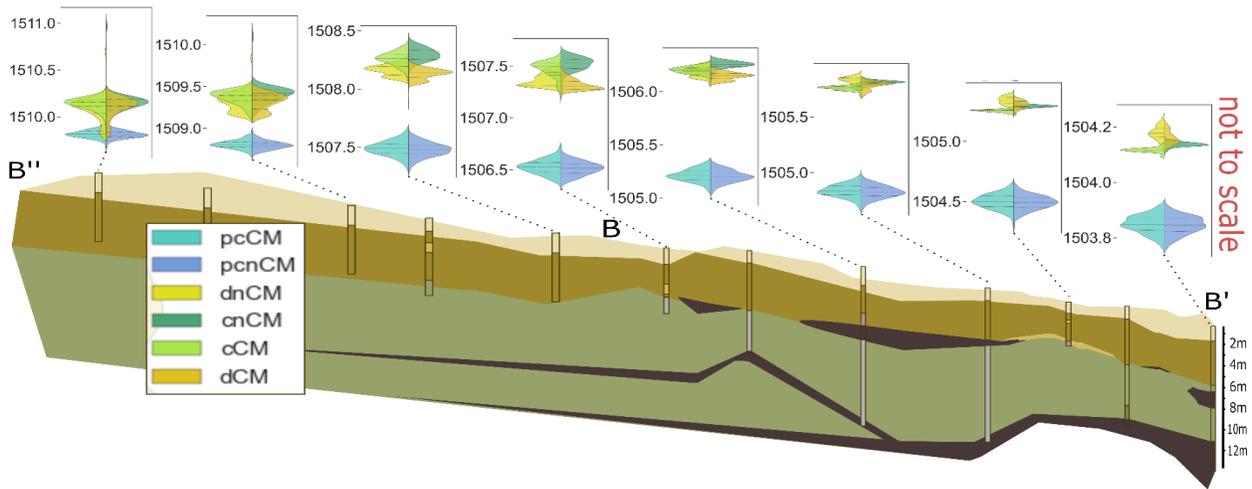
Figure 13: Interpreted stratigraphy from well logs along cross section B' to B and extended north towards the Big Wind River. Simulated water level distributions are shown from the six conceptual models at synthetic upgradient and downgradient wells.

This last section summarizes an interpretation of bimodality in simulated water level distributions predominantly observed in models with the discontinuous clay layer, as seen in Figure 13. The clay layer physically limits and controls where any mixing can occur between aquifers. The bimodal water level response is a reflection to the stochasticity of different possible clay thickness. Specifically, where there are similarities in the peaks of distributions from different CMs, that may indicate many of the models had similar clay structure and simulated similar head responses. Then, some realizations of the discontinuous-clay models may have more clay discontinuity with respect to other dCM realizations or cCM realizations, which therefore allows for more upwelling and mixing to occur during those groundwater simulations. Hence, there is a left-skewed bimodal shape for models with the discontinuous clay structure.

6. Conclusion

This study tested the ability of simulated environmental tracer concentrations to test and reduce conceptual model uncertainty spanned by six alternate conceptual models. The results indicate little sensitivity of hydraulic head and environmental tracer concentration to assess the spatial heterogeneity of recharge; however, these variables are more likely to detect differences in the physical structure of an aquifer or aquifer systems. The first conclusion with tracers and water levels together is that they both identified unique characteristics in kernel density estimates and associated those differences with each alternate conceptual model. The visual and quantitative analysis of normalized simulation data was enhanced by the Wilcoxon Rank-sum test and Kolmogorov-Smirnov test, where there was clear evidence that both tracers and water levels could distinguish between models. In seven useful inter-model Mahalanobis distance calculations, water level simulations were shown to be more similar to one another compared to responses of

simulated tritium and CFC12 concentrations showing there to be more differences between models. SF6 performed worse at reducing uncertainty than water levels in three of those comparisons. With the furthest inter-model distances and the closest match to field data, CFC12 and tritium functioned better than water levels to reduce conceptual model uncertainty of hydrogeologic models.

7. References

Anderson, M. P., Woessner, W. W., & Hunt, R. J. (2015). Modeling Purpose and Conceptual Model. In *Applied Groundwater Modeling*. https://doi.org/10.1016/b978-0-08-091638-5.00002-x

Anderson, M. P., Woessner, W. W., & Hunt, R. J. (2015). Basic Mathematics and the Computer Code. In *Applied Groundwater Modeling*. https://doi.org/10.1016/b978-0-08-091638-5.00003-1

Arnold, B. W., & Gardner, W. P. (2013). Evaluation of Parameter Uncertainty Reduction in Groundwater Flow Modeling Using Multiple Environmental Tracers. *SAND2013-xxxx*. 85000.

Atkinson, A. P., Cartwright, I., Gilfedder, B. S., Hofmann, H., Unland, N. P., Cendón, D. I., & Chisari, R. (2015). A multi-tracer approach to quantifying groundwater inflows to an upland river; assessing the influence of variable groundwater chemistry. *Hydrological Processes*, *29*(1), 1–12. https://doi.org/10.1002/hyp.10122

Balugani, E., Lubczynski, M. W., Reyes-Acosta, L., van der Tol, C., Francés, A. P., & Metselaar, K. (2017). Groundwater and unsaturated zone evaporation and transpiration in a semi-arid open woodland. *Journal of Hydrology*, *547*, 54–66. https://doi.org/10.1016/j.jhydrol.2017.01.042

Batlle-Aguilar, J., Harrington, G. A., Leblanc, M., Welch, C., & Cook, P. G. (2014). Chemistry of groundwater discharge inferred from longitudinal river sampling. *Water Resources Research*, 50(2), 1550–1568. https://doi.org/10.1002/2013WR013591

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, *67*(339), 687–690. https://doi.org/10.1080/01621459.1972.10481279

Boucher, A. (2009) Considering complex training images with search tree partitioning. Computational Geosciences. 35:1151–1158

Bredehoeft, J. (2005). The conceptualization model problem - Surprise. *Hydrogeology Journal*, *13*(1), 37–46. https://doi.org/10.1007/s10040-004-0430-5

Cartwright, I., Cendón, D., Currell, M., & Meredith, K. (2017). A review of radioactive isotopes and other residence time tracers in understanding groundwater recharge: Possibilities,

challenges, and limitations. *Journal of Hydrology*, *555*, 797–811. https://doi.org/10.1016/j.jhydrol.2017.10.053

Castro, M. C., & Goblet, P. (2003). Calibration of regional groundwater flow models: Working toward a better understanding of site-specific systems. *Water Resources Research*, *39*(6), 1–25. https://doi.org/10.1029/2002WR001653

Chambers, L. A., Gooddy, D. C., & Binley, A. M. (2019). Use and application of CFC-11, CFC-12, CFC-113 and SF6 as environmental tracers of groundwater residence time: A review. *Geoscience Frontiers*, *10*(5), 1643–1652. https://doi.org/10.1016/j.gsf.2018.02.017

Cook, P.G., Herczeg, A.L., 2000. Environmental Tracers in Subsurface Hydrology. Springer.

Cressie, N., 1991, Statistics for Spatial Data: John Wiley and Sons, New York, 900 p.

Daddow, R. L. (1996). Water resources of the wind river indian reservation, Wyoming (Vol. 95, No. 4223). US Department of the Interior, US Geological Survey.

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

Department of Energy. 1987a. *Environmental Assessment for Remedial Action at the Riverton Uranium Mill Tailing Site, Riverton, Wyoming,* DOE/EA-0254, prepared by the U.S Department of Energy, UMTRA Project Office, Albuquerque Operations Office, Albuquerque, New Mexico

Department of Energy. (1998a). Field site observational work plan for the UMTRA project site at Riverton, Wyoming (Tech. Rep. No. February).

Department of Energy Information, S., & Book-ope, H. I. (1978). *Fact Sheet Riverton, Wyoming, Processing Site*.

Deutsch, C. V., & Wang, L. (1996). Hierarchical object-based stochastic modeling of fluvial reservoirs. *Mathematical Geology*, *28*(7), 857–880. https://doi.org/10.1007/BF02066005

Doble, R. C., & Crosbie, R. S. (2017). Review: Current and emerging methods for catchment-scale modelling of recharge and evapotranspiration from shallow groundwater. *Hydrogeology Journal*, *25*(1), 3–23. https://doi.org/10.1007/s10040-016-1470-3

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

Dowd, P. (2018). Quantifying the impacts of uncertainty. In *Handbook of Mathematical Geosciences: Fifty Years of IAMG*. Springer International Publishing. https://doi.org/10.1007/978-3-319-78999-6

Ekwurzel, B., Schlosser, P., Jr, W. M. S., Weppernig, R., & Stute, M. (1994). Dating of shallow groundwater: Comparison of thr transient tracers 3H/3He, chlorofluorocarbons, and 85Kr. *Water Resources Research*, *30*(6), 1693–1708.

Elshall, A. S., Tsai, F. T. C., & Hanor, J. S. (2013). Indicator geostatistics for reconstructing Baton Rouge aquifer-fault hydrostratigraphy, Louisiana, USA. *Hydrogeology Journal*, *21*(8), 1731–1747. https://doi.org/10.1007/s10040-013-1037-5

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

Fogg, G. E. (1996). Transition Probability-Based Indicator Geostatistics. *Mathematical Geology*, *28*(4), 453–476. https://doi.org/10.1007/bf02083656

Gardner, W. P., Hammond, G., & Lichtner, P. (2015). High Performance Simulation of Environmental Tracers in Heterogeneous Domains. *Groundwater*, *53*(S1), 71–80. https://doi.org/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. https://doi.org/10.1029/2010GC003353

Gardner, W.P. (2018). Project Proposal: *Reducing Uncertainty in Radionuclide Transport Prediction Using Multiple Environmental Tracers*.

Genereux, D. P., & Hemond, H. F. (1990). Naturally Occurring Radon 222 as a Tracer for Streamflow Generation: Steady State Methodology and Field Example. *Water Resources Research*, *26*(12), 3065–3075. https://doi.org/10.1029/WR026i012p03065

Goble, D. (2018). *Estimating groundwater inflow and age characteristics in an alluvial aquifer along the Little Wind River, Wyoming*.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8), 1–16. https://doi.org/10.1029/2011WR011044

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

Harbaugh AW (2005) MODFLOW-2005, the US Geological Survey modular ground-water model: the ground-water flow process. US Geological Survey, Reston, VA.

He, X. L., Sonnenborg, T. O., Jørgensen, F., & Jensen, K. H. (2014). The effect of training image and secondary data integration with multiple-point geostatistics in groundwater

modelling. *Hydrology and Earth System Sciences*, *18*(8), 2943–2954. https://doi.org/10.5194/hess-18-2943-2014

Healy, R. W. (2010). Estimating groundwater recharge. Cambridge university press.

Højberg, A.L. & Refsgaard, J.C. (2005). Model Uncertainty – parameter uncertainty versus conceptual models. *Water Science and Technology*, *52*(6), 177–186.

Høyer, A.-S., Vignoli, G., Mejer Hansen, T., Vu, L. T., Keefer, D. A., & Jørgensen, F. (2016). Multiple-point statistical simulation for hydrogeological models: 3D training image development and conditioning strategies. *Hydrology and Earth System Sciences Discussions*, *November*, 1–29. https://doi.org/10.5194/hess-2016-567

Houben, P. (2007). Geomorphological facies reconstruction of Late Quaternary alluvia by the application of fluvial architecture concepts. *Geomorphology*, *86*(1–2), 94–114. https://doi.org/10.1016/j.geomorph.2006.08.008

Hu, Y., Moiwo, J. P., Yang, Y., Han, S., & Yang, Y. (2010). Agricultural water-saving and sustainable groundwater management in Shijiazhuang Irrigation District, North China Plain. *Journal of Hydrology*, *393*(3–4), 219–232. https://doi.org/10.1016/j.jhydrol.2010.08.017

Hu, W., Wang, Y. Q., Li, H. J., Huang, M. B., Hou, M. T., Li, Z., She, D. L., & Si, B. C. (2019). Dominant role of climate in determining spatio-temporal distribution of potential groundwater recharge at a regional scale. *Journal of Hydrology*, *578*(August). https://doi.org/10.1016/j.jhydrol.2019.124042

Johnson, R. H., Dam, W. L., Campbell, S., Noël, V., Bone, S. E., Bargar, J. R., & Dayvault, J. (2016). Persistent Secondary Contaminant Sources at a Former Uranium. *Proceedings IMWA 2016, Freiberg/Germany*, *2097*(Doe 1998), 398–404.

Journel, A. G. (2002). Combining knowledge from diverse sources: An alternative to traditional data independence hypotheses. *Mathematical Geology*, *34*(5), 573–596. https://doi.org/10.1023/A:1016047012594

Keefer W.R. (1970). Structural Geology of the Wind River Basin, Wyoming. *US Geological Survey Professional Paper*.

Kikuchi, C. P., Ferre, T. P. A., & Vrugt, J. A. (2015). On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models. *Journal of the American Water Resources Association*, *5*(3), 2–2. https://doi.org/10.1111/j.1752-1688.1969.tb04897.x

King, D. B., & Saltzman, E. S. (1995). Measurement of the diffusion coefficient of sulfur hexafluoride in water. *Journal of Geophysical Research*, *100*(C4), 7083–7088. https://doi.org/10.1029/94JC03313

Koltermann, C. E., & Gorelick, S. M. (1996). Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches. *Water Resources Research*, *32*(9), 2617–2658. https://doi.org/10.1029/96WR00025

Love, J.D and Christiansen, A.C., 1985, Geologic Map of Wyoming: U.S. Geological Survey, scale 1:500,000

Maloszewski, P., & Zuber, A. (1982). Determining the turnover time of groundwater systems with the aid of environmental tracers. *Journal of Hydrology*, 57(3), 207–231. https://doi.org/10.1016/0022-1694(82)90147-0

Manning, A. H., & Solomon, D. K. (2005). An integrated environmental tracer approach to characterizing groundwater circulation in a mountain block. *Water Resources Research*, *41*(12), 1–18. https://doi.org/10.1029/2005WR004178

Narasimhan TN, White AF, Tokunaga T (1986) Groundwater contamination from an inactive uranium mill tailings pile: 2. Application of a dynamic mixing model. Water Resour Res 22(13):1820–1834

Neuman, S. P., & Wierenga, P. J. (2003). A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites. *U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research*, 311.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences 30 (7), 683-691

Remy, N., Boucher, A., & Wu, J. (2009). Applied Geostatistics with SGeMS: A user's guide. *Applied Geostatistics with SGeMS: A User's Guide*, *9780521514149*, 1–264. https://doi.org/10.1017/CBO9781139150019

Remy, N., Boucher, A., & Wu, J. (2011). Variogram computation and modeling. *Applied Geostatistics with SGeMS*, 90–100. https://doi.org/10.1017/cbo9781139150019.007

Richter, H. R. (1981). Occurrence and characteristics of groundwater in the Wind River Basin, Wyoming. *Wyoming Water Resources Data System*, *IV-A*(307).

Rogers, a S. (1958). Physical behavior and geologic control of radon in mountain streams. *In Experimental and Theoretical Geophysics, U.S. Geological Survey Bulletin 1052-E*, 187–212.

Selroos, J. O., Walker, D. D., Ström, A., Gylling, B., & Follin, S. (2002). Comparison of alternative modelling approaches for groundwater flow in fractured rock. *Journal of Hydrology*, *257*(1–4), 174–188. https://doi.org/10.1016/S0022-1694(01)00551-0

Simic, A., Fernandes, R., & Wang, S. (2014). Assessing the Impact of Leaf Area Index on Evapotranspiration and Groundwater Recharge across a Shallow Water Region for Diverse Land Cover and Soil Properties. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *7*, 4881–4884. https://doi.org/10.1109/igarss.2004.1370257

Sinclair, W. J., & Granger, W. (1911). Eocene and Oligocene of the Wind River and Bighorn Basins. *Bulletin American Museum of Natural History.*

Smerdon, B. D., & Gardner, W. P. (2021). Characterizing groundwater flow paths in an undeveloped region through synoptic river sampling for environmental tracers. *Hydrological Processes*, *April 2021*. https://doi.org/10.1002/hyp.14464

Smerdon, B. D., Smith, L. A., Harrington, G. A., Gardner, W. P., Piane, C. D., & Sarout, J. (2014). Estimation des propriétés hydrauliques d'un aquitard, à partir de mesures in-situ de la pression interstitielle. *Hydrogeology Journal*, *22*(8), 1875–1887. https://doi.org/10.1007/s10040-014-1161-x

Strebelle, S. 2002. Conditional simulation of complex geo- logical structures using multiple-point statistics. Mathematical Geology 34, no. 1: 1–21.

Suzuki, S., Caumon, G., & Caers, J. (2008). Dynamic data integration for structural modeling: Model screening approach using a distance-based model parameterization. *Computational Geosciences*, *12*(1), 105–119. https://doi.org/10.1007/s10596-007-9063-9

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

Tóth, J. (1962). A Theory of Groundwater Motion in Small Drainage Basins. *Journal of Geophysical Research*, *67*(11). http://www.agu.org/pubs/crossref/1962/JZ067i011p04375.shtml

Troldborg, L., Refsgaard, J. C., Jensen, K. H., & Engesgaard, P. (2007). The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal*, *15*(5), 843–860. https://doi.org/10.1007/s10040-007-0192-y

Troldborg, M., Nowak, W., Tuxen, N., Bjerg, P. L., Helmig, R., & Binning, P. J. (2010). Uncertainty evaluation of mass discharge estimates from a contaminated site using a fully Bayesian framework. *Water Resources Research*, *46*(1), 1–19. https://doi.org/10.1029/2010WR009227

Turnadge, C., & Smerdon, B. D. (2014). A review of methods for modelling environmental tracers in groundwater: Advantages of tracer concentration simulation. *Journal of Hydrology*, *519*(PD), 3674–3689. https://doi.org/10.1016/j.jhydrol.2014.10.056

Wang S, Shao J, Song X, Zhang Y, Huo Z, Zhou X (2008) Application of MODFLOW and geographic information system to groundwater flow simulation in North China Plain, China. Environmental Geology 55: 1449–1462

Wanninkhof, R., Mulholland, P. J., & Elwood, J. W. (1990). Gas exchange rates for a first-order stream determined with deliberate and natural tracers. *Water Resources Research*, *26*(7), 1621–1630. https://doi.org/10.1029/WR026i007p01621

White, A. F., Delany, J. M., Narasimhan, T. N., & Smith, A. (1984). *Groundwater contamination from an inactive uranium mill tailing pile 1. application of a chemical mixing model*. *20*(11), 1743–1752.

Ye, M., Neuman, S. P., & Meyer, P. D. (2004). Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research*, *40*(5), 1–17. https://doi.org/10.1029/2003WR002557

Zheng, C. and Wang, P.P., 1998. MT3DMS, A Modular Three-Dimensional Multispecies Transport Model for Simulation of Advection, Dispersion and Chemical Reactions of Contaminants in Groundwater Systems, Documentation and User's Guide. Departments of Geology and Mathematics, University of Alabama.

Zheng, M., Bruyn, W. J. De, & Saltzman, E. S. (1998). *Measurements of the diffusion coefficients of CFC-11 and CFC-12 in pure water and seawater*. *103*, 1375–1379.