water

MDPI

Article

Validating a Calibrated Model of a Groundwater Pump-And-Treat System Using Robust Multiple Regression

Michael Rush *, Leslie Gains-Germain, Lauren M. Foster and Tom Stockton

Neptune and Company, Inc., 1435 Garrison St, Lakewood, CO 80215, USA
* Correspondence: mrush@neptuneinc.org

Abstract: Validation of groundwater models is relatively challenging due to the need to reserve scarce water level data for calibration targets. In addition, traditional statistical validation metrics are unintuitive for non-technical audiences and do not directly identify model behaviors that require further refinement. We developed a novel model validation method that analyzes rate change events at pump-and-treat wells and statistically compares the water level responses at nearby monitoring wells between the data and model. The method takes advantage of events that occur alongside ambient pumping, unlike parameter estimation techniques that require independent drawdown or recovery events. The ability of the model to match well connections that are evident (or not evident) in the observations is characterized statistically, leading to four decision scenarios: model matches the observed connection (1) or lack thereof (2), model exhibits a connection that is not observed (3), or model over- or understates the observed connection (4). The method is applied to an FEHM-based groundwater flow and transport model that is shown to match 84.5% of the well connections analyzed. The method provides novel perspectives on the influence of calibration targets on the flow field and suggests that although the overall effect of drawdown targets was to improve the model, the choice of target well pairs creates flow pathways that may be inconsequential during normal operational conditions. The model adequately matches the flow over short spatial scales (<800 m) and over-represents the flow over greater distances (300–1200 m), suggesting the need for "null" drawdown targets in subsequent rounds of calibration.

Keywords: validation; multiple regression; statistical tools; contaminant transport; groundwater models

check for updates

Citation: Rush, M.; Gains-Germain, L.; Foster, L.M.; Stockton, T. Validating a Calibrated Model of a Groundwater Pump-And-Treat System Using Robust Multiple Regression. *Water* **2024**, *16*, 3178. https://doi.org/10.3390/w16223178

Academic Editor: Dimitrios E. Alexakis

Received: 26 September 2024
Revised: 18 October 2024
Accepted: 25 October 2024
Published: 6 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Model validation generally involves comparing model predictions to a set of data that have not been used in the development of the model and is relatively challenging for groundwater models due to the expense associated with obtaining subsurface observations. In the context of groundwater contaminant remediation, limited subsurface data can lead to regulatory decisions with significant environmental and public health consequences being made without complete information. Whereas significant effort has been devoted to groundwater model calibration schemes and uncertainty quantification in recent years, similar method development effort has not been applied to formalize model validation since the 1990s and early 2000s. As the ability of groundwater models to reproduce data that have not been targeted during calibration is a key measure of their predictive value, there is a strong need to develop model validation approaches that take advantage of relatively scarce available data.

In the context of environmental modeling, validation consists of a demonstration that a model suitably represents site-specific features and behaviors of the underlying system and makes predictions that are sufficiently accurate for decision-making. It is often paired with the term "verification", which refers to a demonstration that a general set of equations, modeling code, and numerical solution scheme matches existing analytical solutions and

laboratory tests, as well as "calibration", which refers to the process of adjusting parameter values to minimize the difference between model predictions and target observations.

Significant debate has occurred over the years surrounding the purpose, feasibility, and limits of model validation in broader discussions of the philosophy of science, hypothesis testing, and the appropriate use of environmental models in decision-making, particularly in the high-level radioactive waste and contaminant transport communities [1–10]. Perhaps no one has so concisely summarized the philosophical limits of environmental models as [11]: "a model cannot predict what will happen in the future. However, following proper uncertainty analysis, it may be able to predict what will NOT happen in the future". Of course, no environmental model can be proven as correct due to such limitations as parameter uncertainty and incompleteness, differences in scale between system property measurements and the model grid, solution non-uniqueness, simplifying assumptions, and the inherent variability of natural systems [6,11]. In the context of groundwater contaminant transport models, validation cannot guarantee predictions of future contaminant locations with high accuracy, due to incomplete knowledge of the subsurface, or with high precision, due to the dispersive nature of contaminant migration and solution non-uniqueness. In contrast to groundwater flow models, for which good agreement between simulations and observations in a data-rich region of the subsurface often implies good agreement in unconstrained locations due to the speed of pressure wave propagation, contaminant transport models that match the data in high-concentration locations have much more limited predictive ability outside of the current plume boundaries.

Considering these limitations, validation serves as an iterative process of "confidence building" that rigorously evaluates model predictions, provides "reasonable assurance" that model results are acceptable, rejects inadequate models, and identifies model flaws that can be addressed through subsequent iterations of calibration and reconceptualization [2–6]. The modeler must "demonstrate the degree of correspondence between the model and the material world it seeks to represent and to delineate the limits of that correspondence" [6]. Debate aside, the modeling community has agreed that validation should always be conducted using data that were not directly used in model calibration [1–5,7,12,13].

Various model validation strategies have been proposed, such as the "post-audit" [7], in which the ability of a model to predict the future is tested using field data collected following the completion of model development, the "split-sample technique", in which some observations are excluded from calibration and reserved for validation, that has become ubiquitous in recent years [12,14–17], and K-fold cross-validation, in which the predictions of statistical learning models trained using many subsets or partitions (often 10) of the overall data are compared to the remaining validation data [18–20]. However, many groundwater models are developed for deep subsurface systems where observations are difficult and expensive to obtain, so reserving data for validation is generally not cost-effective [2–4,7], and as a result, much more emphasis has been placed in recent years on uncertainty quantification and sensitivity analysis that transparently demonstrates the dependence of model predictions on key parameters [11,21–24]. Thus, there is a strong need for innovative methods that forage for independent validation opportunities amidst scarce observations.

Statistical techniques have long played a major role in environmental model validation [8,25] due to the objective nature of statistical measures that provide a quantifiable degree of confidence to regulators and stakeholders. Most model validation is accomplished using the coefficient of determination [8,12,16], regression coefficient [8,16], the root-mean-square error (RMSE) or mean absolute error (MAE) [8,17], the Nash–Sutcliffe Efficiency Coefficient [15,17,18], percent bias [18], or other measures such as predictive skill and likelihood ratio [19]. These methods typically present scatter plots of simulated values against observed, with the statistical measure describing the correlation and magnitude of model error, without identifying the mechanisms leading to data-model disagreement or identifying specific areas of improvement for the model. Given a threshold for determining the accuracy of the model (e.g., RMSE < 10%) or a desired decision confidence

level (e.g., 90%), one can objectively evaluate the suitability of an environmental model for decision-making. However, existing statistical approaches have the key disadvantage that such metrics are not particularly intuitive for non-technical audiences, do not spatially describe model error, and do not identify model features, behaviors, or mechanisms that require further development and refinement, presenting a need for novel approaches to validation that both appeal to regulators and directly inform further development.

Groundwater contaminant transport models benefit from a strong underlying flow model that simulates the response of an aquifer to "stresses" that can include pumping, injection, stream recharge, recharge derived from infiltration, evapotranspiration, among others. These stresses are typically prescribed to the model as inputs, and the ability of the model to reproduce the influence of these stresses is assessed by comparing simulated hydraulic head outputs to water level observations at monitoring wells. During calibration, aquifer properties such as hydraulic conductivity and storativity are iteratively adjusted to obtain the best agreement between simulated and observed heads.

A classic example is a drawdown at a monitoring well in response to a pumping well. Groundwater model calibration often seeks to match data obtained from an aquifer test, in which an extraction well is pumped at a known discharge rate and water levels at a nearby monitoring well or piezometer are recorded at regular intervals. Given many such tests and associated observations, some might be used for calibration, with others reserved for validation. However, aquifer tests are performed relatively infrequently, so these data are almost always included in the calibration and cannot be used for validation. Any opportunistic instances in the pumping record during which extraction commences at a single well in the absence of other pumping are also generally reserved for calibration targets because estimating hydraulic flow parameters from pumping events requires such conditions.

Ideally, a validation exercise should employ data and methods that differ significantly from those underlying the development of the model and should make minimal assumptions about the behavior of the system. From the perspective of contaminant transport modeling, it is important for the underlying flow model to match both the monitoring well responses that are suggested by the data and the absence of responses in the data. In other words, a model must reproduce some well connections and limit the connections between other wells; subsurface flow should not be simulated between wells that are hydrogeologically disconnected. Thus, any validation approach should be able to evaluate both types of model capabilities.

In pump-and-treat systems, which treat groundwater contaminant plumes using an array of wells that extract contaminated water and then inject treated water back into the subsurface, wells are often turned on or off for a variety of operational or institutional reasons. However, such events almost always occur against the backdrop of other pumping, which complicates efforts to isolate the effects of a single pumping well on nearby monitoring wells. In response to this problem, Harp and Vesselinov [26] presented a successful approach in which a superposition of Theis solutions was nested within a calibration algorithm [23] to "fingerprint" transient water level responses to changes in pumping rates at several water supply wells. Theis [27] curves belong to a family of empirical solutions that can be used to fit transient water level drawdown in confined aquifers, with parameters S (storativity) and T (transmissivity), when several assumptions are met: constant pumping rate, horizontal flow, homogeneous aquifer properties (S and T), and uniform thickness (often denoted "b").

The authors also included a linear drawdown term to account for background temporal trends in water levels that could not be explained by pumping. Interestingly, the Theis-based approach can determine when a monitoring well is hydrogeologically disconnected from a pumping well: in these cases, the calibration algorithm repeatedly converged on S and T values that led to negligible drawdown at the monitoring well, indicating a geologic barrier to pressure wave propagation. Accordingly, this physics-based method can broadly be applied to the development and calibration of groundwater models simulating confined aquifer flow; however, as the method makes several assumptions about the behavior of the

system, it is not as attractive for validation. Whereas calibration is intended to implement a specific conceptual model (e.g., confined flow), validation methods should retain the capability to identify the conceptual model error (e.g., due to the presence of unconfined or leaky confined behavior) by avoiding such assumptions.

In this study, we present a statistical validation scheme in which many pumping events (when a well completely turns on or off) spanning a long period of record (2016–2022) for pumping well/monitoring well pairs are analyzed to isolate the influence of a single well from other pumping and background trends. Our generalized approach makes no assumptions about the underlying nature of the aquifer system (confined, two-dimensional flow, homogeneous, etc.), background water level trends (linear), or the transient response of the system (Theis solution). Instead, the method is based on the simple assumption that extraction leads to water level declines, and a pause in extraction leads to a corresponding recovery in water levels (and vice versa for injection). The tradeoff of such a generalized approach is that greater care must be taken when applying the method and defining the time periods over which water levels will be compared, which will depend on local site characteristics.

Using multiple linear regression, we analyze the well connections (or lack of connection) evident in the data and then determine whether the model reproduces those connections or lack thereof. Our approach can broadly be applied to any groundwater model (confined, unconfined, leaky confined, etc.) or any system in which a positive or negative stress consistently leads to a positive or negative response. In contrast to traditional validation approaches that simply characterize the agreement between simulations and observations using a statistical measure, our method identifies mechanisms in the model that require further refinement, regions of the subsurface that are inadequately represented in the model, provides nuanced spatiotemporal model performance information, and generally suggests opportunities for model improvement. Whereas cross-validation approaches [18–20] evaluate model performance against partitions of the data that have been sequentially omitted, our approach employs an entirely new set of data that has not been used in the development of the model.

2. Materials and Methods

2.1. Site and Model Overview

A plume of hexavalent chromium (referred to as Cr in this document) with concentrations exceeding the 50 parts per billion (ppb) New Mexico groundwater standard was detected in multiple monitoring wells beneath Los Alamos National Laboratory (LANL). The highest recorded concentration in these wells is in the order of 1000 ppb. The primary source of Cr is effluent containing between 31,000 and 72,000 kg of Cr that was discharged by a LANL power plant into Sandia Canyon between 1956 and 1972. Since Cr concentrations were discovered in 2005, extensive investigations into migration of the plume have been ongoing [28–36]. A pump-and-treat system known as the "Interim Measure" or IM network, consisting of five injection wells (CrIN) and five extraction wells (CrEX), was developed to accomplish two key objectives: (1) achieve and maintain the 50-ppb downgradient Chromium plume edge within the laboratory boundary and (2) hydraulically control plume migration in the eastern downgradient portion of the plume [31,32].

Aquifer tests at the site have led to a conceptual site model that considers the upper portions of the aquifer to be unconfined. Confined/leaky-confined behavior increases with increasing depth in a heterogeneous, non-continuous manner due to high anisotropy from interbedded alluvial deposits [30,33,34,36–38]. Such spatial variability in flow system dynamics (unconfined, confined, leaky confined) presents major challenges for the interpretation of aquifer tests, as a variety of well functions [27,39,40] could be applied with different results. The general direction of groundwater flow is to the east/southeast, and the water table in the vicinity of the Cr plume is significantly (10–100 times) flatter than the surrounding areas.

An FEHM-based numerical fate and transport model of the Cr plume (the Chromium Model, or CM) was developed to simulate the migration of the plume, optimize operations of the IM, and map capture and flood zones for IM wells [35]. The model was calibrated using the Model Analysis and Decision Support (MADS) software, with targets of Cr concentrations, hydraulic head, hydraulic head gradients, and drawdown responses for 40 well screens. The key calibration parameters were Cr source concentration and extent, hydraulic conductivity, specific storage, porosity, dispersivity, recharge rates, and elevation of hydraulic head boundaries. The confined/leaky-confined behavior that prevails with increasing depth in a heterogeneous manner was represented in the model through pilot points that allowed for vertical hydraulic conductivity to vary with depth. In the development of drawdown targets, three opportunistic periods in the pumping record ranging from 1 to 15 days were identified in which a single well began pumping when no other Interim Measure wells were active and thus the drawdown response at nearby monitoring wells could be attributed to that pumping well. These rare periods were used to develop drawdown targets for the calibration of hydraulic conductivity and specific storage. The calibration process led to four deterministic calibrations (calibrations 1–4) representing unique flow pathways and source areas that were identified as part of a range of possible solutions, all of which align with the conceptual site model and closely match calibration targets.

Following the calibration process, validation was required to demonstrate the ability of the four model calibrations to reproduce water level responses to pumping that were not explicitly included in the calibration as targets. As the exceedingly rare instances where pumping was initiated at a single well during a period when no other wells were active had already been incorporated in the calibration, the validation process must make use of other events. In subsequent sections, we describe the validation scheme that was devised to statistically analyze the many events during which pumping was initiated or ceased while other pumping wells were active. The method consists of event selection, response characterization, regression analysis of events for well pairs, and well-pair classification. The criteria used in each step of the analysis (e.g., minimum sample size, significance level, minimum and maximum event durations) were selected based on a thorough review of individual events and well-pair classifications by an interdisciplinary team of hydrogeologists and statisticians. Through this process, simple filters, checks, parameters, and strategies were pragmatically devised to address common issues affecting interpretation of water level responses. Examples of the pumping events and water level responses in the data and model calibrations are provided in Figures 1 and 2.

2.2. Event Selection

The full period of record (2015 to 2022) for each of the ten IM extraction and injection wells is analyzed to identify instances where a pump turned completely on or off and the pumping rate remained constant for at least two days. See Figures 1 and 2 for examples of pumping events. If the pumping rate remains constant for longer than seven days, only the first seven days are retained as part of the event. The two-day minimum was selected to allow sufficient time for water levels to progress past the rapid initial response phase and begin their asymptotic approach to steady state (see Figures 1 and 2). The seven-day maximum was selected to minimize the influence of long-term water level trends (e.g., due to recharge) on results. These two thresholds are key characteristics of the method that are expected to vary based on site-specific characteristics and were selected based on a thorough review of many water level responses. With time periods shorter than two days, monitoring wells are still rapidly adjusting to pumping rate change events, whereas with time periods longer than seven days, long-term trends start to interfere with the interpretation of water levels. Each event is defined by the time associated with the pumping rate change (termed t_{start}) and the time associated with the next pumping rate change, or seven days following t_{start}, whichever is earlier (termed t_{end}).

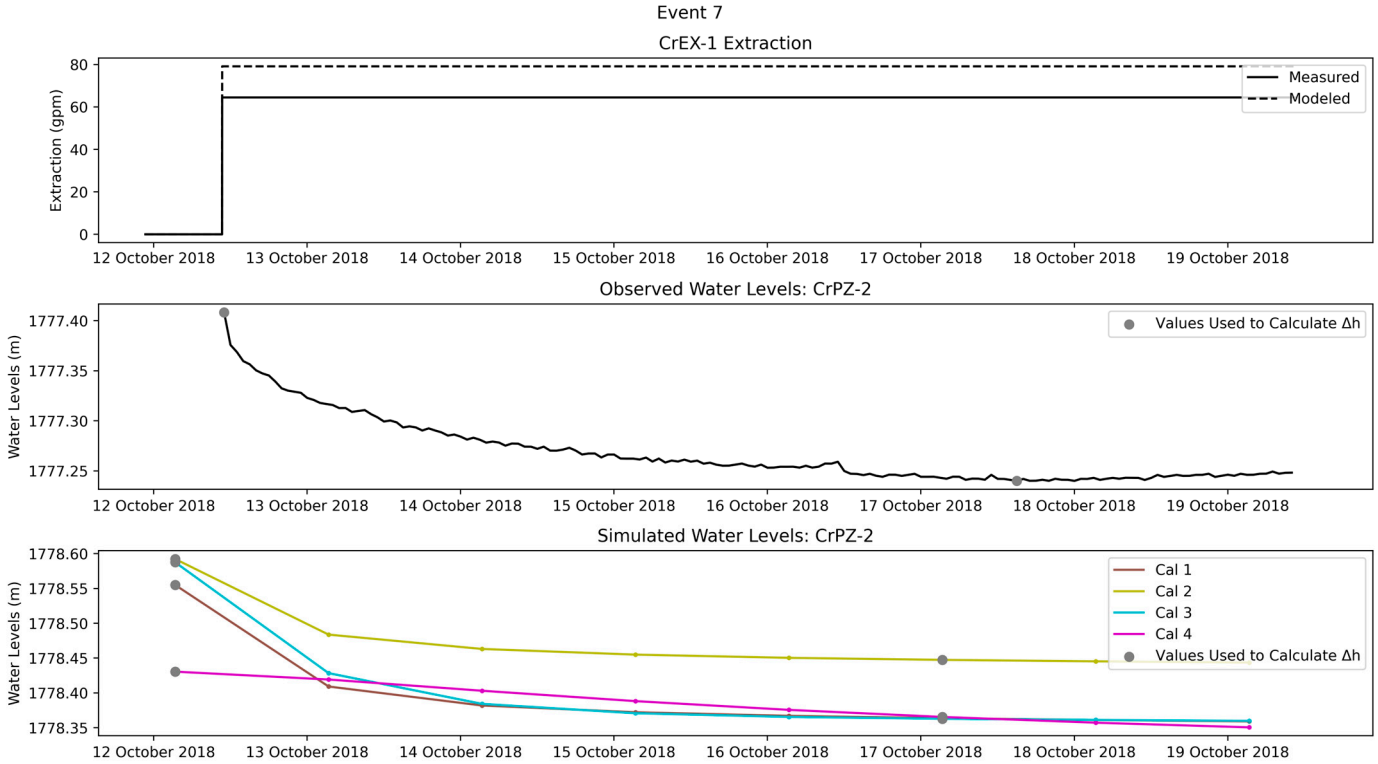

Figure 1. Example of CrEX-1 pumping event (**top**), corresponding CrPZ-2 monitoring well drawdown response in the data (**middle**) and in multiple model calibrations (**bottom**).

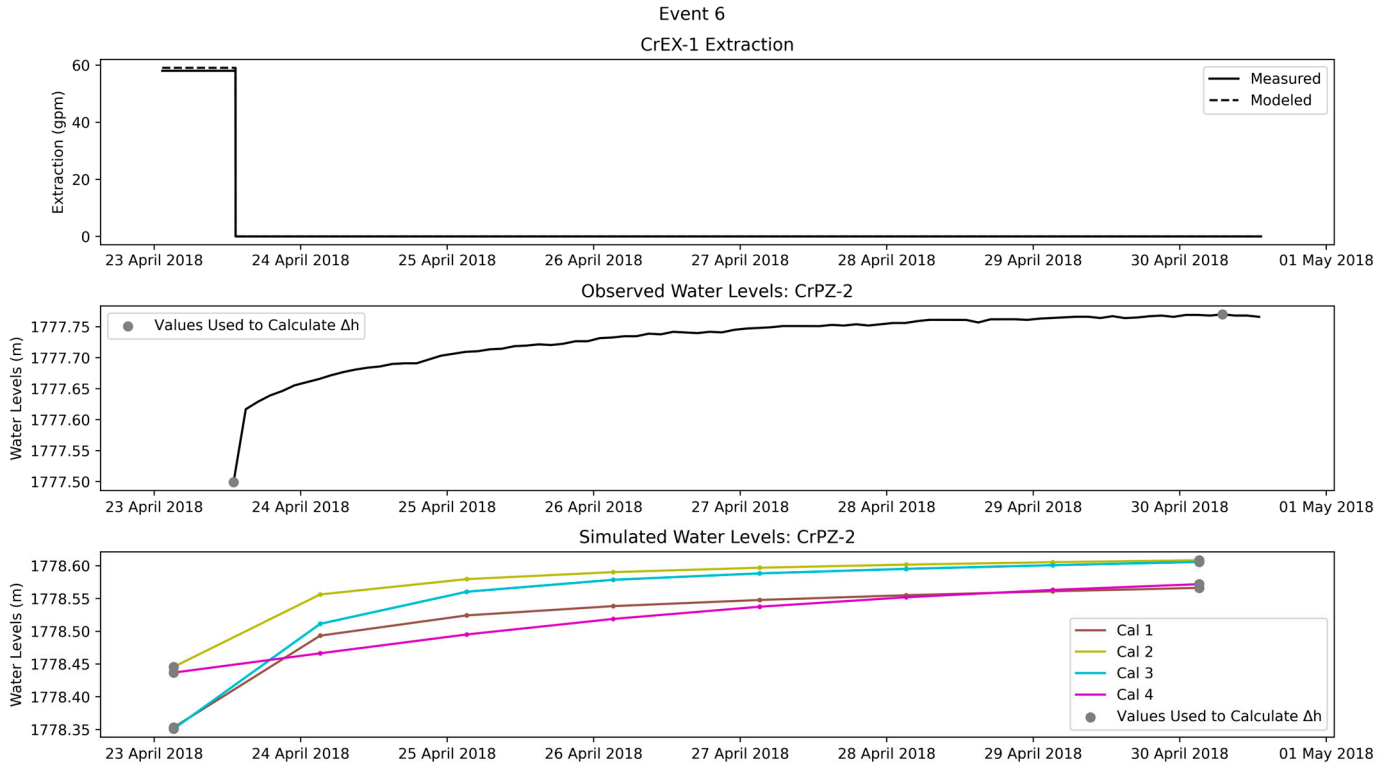

Figure 2. Example of CrEX-1 pumping event (**top**), CrPZ-2 monitoring well water level recovery response in the data (**middle**) and multiple model calibrations (**bottom**).

Next, pumping rate change events which occur in the data record but are not represented in the model are excluded (Table 1, Column 3). The pumping record was smoothed

for the CM to improve numerical efficiency and stability by grouping pumping and recovery periods together with a weighted average (a common practice in groundwater modeling). Pumping events with duration less than 96 h or a recovery less than 12 h are grouped together in the weighted average because the influence of such short-term changes in pumping rate on groundwater flows and contaminant concentrations can be adequately represented using a weighted average. As the model cannot be expected to reproduce water level responses to pumping rate change events that are prescribed to the model in a temporally averaged fashion, the instances in which the modeled pumping scheme differed from the observed record are not included in the validation (25–60%, see Table 1). Specifically, any cases in which the well was active (inactive) in the data record, but inactive (active) in the model due to smoothing are removed, leaving some cases in which the pumping rate differs slightly due to the weighted averaging scheme (e.g., Figure 1). Excluding such events limits the ability of the validation to characterize the model's representation of short-term transient fluctuations; however, such events are inconsequential for broader groundwater flow and contaminant trends, and sufficient pumping events remain in the validation for a rigorous analysis.

Finally, any extraction or injection well with less than 15 identified events is excluded from the validation due to insufficient sample sizes (CrEX-4 and -5, CrIN-1 and -2), leaving six wells (CrEX-1 to -3, CrIN-3 to -5). Reducing the sample size threshold below 15 events would lead to unreliable validation results based on a small subset of the pumping record, whereas increasing the sample size threshold would exclude too many extraction wells (CrEX-1 and CrEX-3), severely limit the scope of the validation, and prevent model performance evaluation in those regions of the domain. Although some extraction and injection wells were excluded based on sample sizes (CrEX-4, -5, CrIN-1, -2), this decision did not influence the monitoring wells that were included in the validation, except for a single well pair (CrEX-1, R-35a) that fell below the sample size threshold and was thus excluded from the validation. Other monitoring wells that did not have sufficient sample sizes due to limited periods of record were R-70 S1, R-70 S2, and R-33 S2. However, the remaining 25 well screens provide a thorough representation of the site, both vertically and laterally (see Appendix C Figures A25–A30). Numbers of events analyzed are presented in Table 1.

Table 1. Pumping events identified for validation. Rows shaded green indicate wells that were included in the validation, and rows shaded orange indicate wells excluded from the validation due to low sample sizes (<15 events).

Well	Initial Number of Events Identified	Number of Events Excluded	Percent Excluded	Number of Remaining Validation Events	Included in Validation?
CrEX-1	35	16	46%	19	Yes
CrEX-2	32	8	25%	24	Yes
CrEX-3	24	6	25%	18	Yes
CrEX-4	16	8	50%	8	No
CrEX-5	15	9	60%	6	No
CrIN-1	25	14	56%	11	No
CrIN-2	19	9	47%	10	No
CrIN-3	34	12	35%	22	Yes
CrIN-4	37	15	41%	22	Yes
CrIN-5	38	16	43%	22	Yes

2.3. Response Characterization

Changes in pumping rate Δq at the pumping wells introduced in Section 2.2 are compared to changes in water level at monitoring wells Δ_h. As pumping rates and water levels invariably fall on different scales, the changes in pumping rates and water levels are normalized by their respective standard deviations σ_q and σ_h, similar to a z-score. This approach provides an intuitive definition of pumping rate changes in terms of the number of standard deviations away from the initial rate. While alternate normalization schemes

(e.g., maximum, or half-maximum) were explored, standard deviation provided the most reliable set of relationships between pumping rate changes and water level responses, and provided a meaningful basis for outlier detection, as described later in this section. Changes in pumping rate are computed as:

$$\Delta q = \frac{q_{new} - q_{initial}}{\sigma_q} \quad (1)$$

When the well begins pumping, $q_{initial}$ is zero, q_{new} is the pumping rate, and Δq is positive. When the well is shutting down, $q_{initial}$ is the pumping rate, q_{new} is zero, and Δq is negative. Pumping rates are considered positive for both extraction and injection wells. For a given event, Δq is nearly identical in both the data and the model calibrations, as the pumping rates (q) are prescribed directly to the model as boundary conditions.

Water level responses Δh are computed from the maximum, minimum, and initial water levels during the event in the data:

$$\Delta h = \begin{cases} if \left| h_{max|_{t_{start}}^{t_{end}}} - h_{t_{start}} \right| < \left| h_{min|_{t_{start}}^{t_{end}}} - h_{t_{start}} \right| & \frac{h_{min|_{t_{start}}^{t_{end}}} - h_{t_{start}}}{\sigma_h} \\ else & \frac{h_{max|_{t_{start}}^{t_{end}}} - h_{t_{start}}}{\sigma_h} \end{cases} \quad (2)$$

If h_{max} is closer to the initial water level $h_{t_{start}}$, water levels decrease during the event, Δh is calculated as the difference between h_{min} and $h_{t_{start}}$, and Δh is negative. Otherwise, h_{min} is closer to the initial water level $h_{t_{start}}$, water levels increase during the event, Δh is calculated as the difference between h_{max} and $h_{t_{start}}$, and Δh is positive. Figures 1 and 2 show the values used to calculate Δh for the data.

Analyzing the absolute change in water level over the pumping event timespan reduces a transient drawdown or recovery curve to a single, simplified response metric that can be compared between the data and model and allows any long-term water level trends to be neglected. This approach avoids detailed assumptions about the physics of the system, while conforming to widely accepted analytical solutions for steady-state water level drawdown as linear functions of pumping rate in confined [41] and unconfined [42] aquifers. Although non-linearities arise in groundwater flow for several reasons (e.g., unconfined transient flow to a well, high spatial contrast in subsurface properties), the use of a minimum time threshold (two days) ensures that water levels have progressed past the non-linear phase of the response, based on a review of many water level responses.

Two additional checks are performed during the calculation of Δh for the data to ensure that there is sufficient water level data to conduct the analysis for the given event and well pair and to remove the influence of outliers on the maximum or minimum water level due to transducer slips or other anomalous artifacts of data acquisition. These additional filters were developed by reviewing many events and well pairs in the form of Figures 1 and 2 to identify common issues affecting the validation scheme and devise appropriate fixes. For a given well pair, the data availability check verifies that water level data exists within 24 h of both t_{start} and t_{end}, and excludes the event from the well pair analysis if those conditions are not met. The data availability check reduces the number of events for three of forty monitoring wells below the desired sample size (minimum 15 events), and those wells are accordingly excluded from the analysis. The outlier check compares the water level minimum to the next lowest value for the first case in Equation (2) or compares the water level maximum to the next highest value for the second case of Equation (2). If the difference between the minimum and second lowest, or maximum and second highest exceeds the standard deviation of the water levels during the event, an iteration is performed until the difference between the value and the next lowest/highest value falls below the standard deviation. This value is then used in Equation (2). The standard deviation was selected as a threshold by plotting many clear examples of outliers and transducer slips and comparing the magnitude of the outliers to the true maximum

identified by a group of expert hydrogeologists, and then defining a meaningful threshold for excluding the outliers.

Water level responses Δh for model calibrations are calculated in a similar fashion to the data. However, rather than using h_{max} and h_{min} to compute Δh, as was carried out for the data, the date and time associated with the maximum or minimum water level in the data is used to identify the closest corresponding model output. This approach guarantees that Δh for model outputs and the data correspond to matching time intervals. Figures 1 and 2 show the values used to calculate Δh for the model calibrations.

2.4. Huber Regression

Next, Huber regression [43] is performed for all pumping wells—monitoring well pairs, with the independent variable being the change in pumping rate Δq and the dependent variable being the change in water level Δh. Huber regression is selected to mitigate the effects of outlying observations, which are generally more prevalent in the model than in the data (see additional discussion in Appendix A). The Huber loss function is piecewise and less sensitive to outliers than standard least squares regression [44].

The approach combines two linear models. A simple linear regression with only the observed Δh and Δq (3) evaluates the data response:

$$\begin{gathered} \Delta h_j = \beta_0 + \beta_{data}\Delta q_j + \varepsilon_j \\ \varepsilon \sim N\left(0,\ \sigma_1^2\right) \end{gathered} \tag{3}$$

A multiple linear regression that includes Δq and Δh for both the data and the model calibrations (4), with the data Δh distinguished from the modeled Δh using the categorical variables "data", "cal1", "cal2", "cal3", and "cal4", is used to evaluate the model response:

$$\begin{gathered} \Delta h_j = \beta_0 + \beta_{1_{cal_i}} I_{cal_i} + \beta_2 \Delta q_j + \beta_{3_{cal_i}} \Delta q_j I_{cal_i} + \varepsilon_j \\ I_{cal\,i} = \begin{cases} 0\ for\ cal \neq cal_i \\ 1\ for\ cal = cal_i \end{cases} \\ \varepsilon \sim N\left(0,\ \sigma_1^2\right) \end{gathered} \tag{4}$$

i represents the calibration number, $i \in 1,\ 2,\ 3,\ 4$, and j indicates the event number, $j \in 1, 2,\ 3\ , \ldots N_p$ where N_p is the number of events for well pair P, where $P \in 1,\ 2,\ 3, \ldots 240$. (3) yields a p-value for β_{data} (termed $p_{\beta_{data}}$) which is used to characterize the relationship between the response of the monitoring well and identify significant well connections in the data. (4) fits all Δh for a given well pair with "data" assigned as the reference level or baseline, and yields four slope parameters ($\beta_{3_{cal_1}}$, $\beta_{3_{cal_2}}$, $\beta_{3_{cal_3}}$, and $\beta_{3_{cal_4}}$) that describe the difference between the data and model relationships. The p-values associated with $\beta_{3_{cal_i}}$ (termed $p_{\beta_{3_{cal_i}}}$) are used to characterize the extent to which the modeled response of the monitoring well matches the response in the data. While classifying observed well connections using the $p_{\beta_{data}}$ from (4) rather than (3) was explored, the inclusion of simulated Δh and Δq in (4) unduly influences $p_{\beta_{data}}$. This has the effect of allowing model output to influence the characterization of well connections in the data, which is not a desirable feature for a validation tool that should consistently classify observed well connections, regardless of model output.

2.5. Well-Pair Classification

Next, the five p-values introduced above ($p_{\beta_{data}}$ and $p_{\beta_{3_{cal_i}}}$) are used to characterize the ability of the model calibrations to reproduce monitoring well—pumping well connections that are suggested by the data. As discussed in Section 1, groundwater models should represent well connections observed in the aquifer, and limit connections between other

wells that are hydrogeologically disconnected. The need to evaluate both types of model capabilities naturally gives rise to a set of classifications (Figure 3).

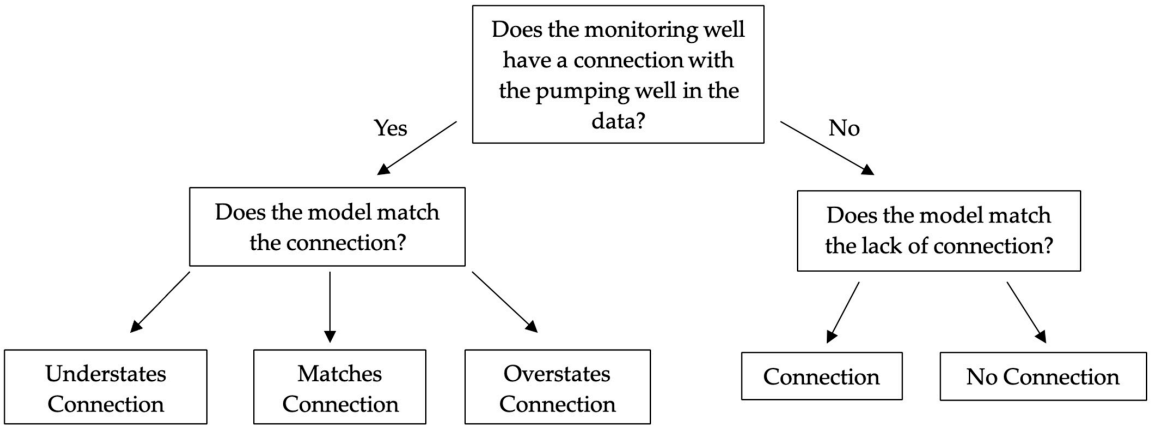

Figure 3. Flow chart describing well-pair classification scheme.

An analysis of *p*-values and slopes is used to classify the modeled and observed responses for each well pair (Table 2). Broadly, the *p*-value associated with the data slope ($p_{\beta_{data}}$) is used to identify significant well connections evident from the data, and the *p*-values associated with the data-model slope difference ($p_{\beta_{3_{cal_i}}}$) are used to determine whether the model responses matches the data response. A significance level (termed α) of 0.01 was selected through a review of individual events and classifications by an interdisciplinary team of hydrogeologists and statisticians, and a sensitivity analysis is provided in Section 3 describing the effects of significance level on results.

Table 2. Slope and *p*-value criteria used to classify well-pair relationships.

Classification	Sign of Data Slope: $\frac{\beta_{data}}{\lvert\beta_{data}\rvert}$	*p*-Value of Data Slope: $p_{\beta_{data}}$	*p*-Value of Data-Model Slope Difference: $p_{\beta_{3_{cal_i}}}$	Magnitude of Model Slope $\lvert\beta_{3_{cal_i}}\rvert$	Sign of Model Slope $\frac{\beta_{3_{cal_i}}}{\lvert\beta_{3_{cal_i}}\rvert}$
Model Understates Connection	$\frac{\beta_{data}}{\lvert\beta_{data}\rvert} < 0$ for extraction wells	$<\alpha$	$<\alpha$	$\lvert\beta_{3_{cal_i}}\rvert < \lvert\beta_{data}\rvert$	$\frac{\beta_{data}}{\lvert\beta_{data}\rvert} = \frac{\beta_{3_{cal_i}}}{\lvert\beta_{3_{cal_i}}\rvert}$
Model Matches Connection		$<\alpha$	$>\alpha$		
Model Overstates Connection		$<\alpha$	$<\alpha$	$\lvert\beta_{3_{cal_i}}\rvert > \lvert\beta_{data}\rvert$	
Model Matches No Connection	$\frac{\beta_{data}}{\lvert\beta_{data}\rvert} > 0$ for injection wells	$>\alpha$	$>\alpha$		
Model Overly Connected		$>\alpha$	$<\alpha$		

In addition to its *p*-value, the data slope β_{data} itself is used to distinguish between responses to extraction versus injection wells. In a pump-and-treat system like the Chromium Interim Measure, extraction and injection rates are typically dependent on one another because extracted water is subsequently treated and then re-injected. Thus, a reduction in extraction rate leads to a consequent reduction in injection rate. In examining relationships between monitoring wells and extraction well pairs, then, it is not uncommon to observe an increase in water level at the monitoring well associated with an increase in pumping, due to the consequent increase in injection at a nearby injection well. While, in principle, this complicated relationship presents challenges for sorting out the counteracting effects of extraction and injection wells, the statistical approach neatly clarifies such ambiguity by imposing additional criteria on the well pair relationships deemed significant in the data. In classifying monitoring well—extraction well pairs, only those relationships with a negative slope (i.e., increase in pumping leads to decrease in water level) are retained. Similarly, for monitoring well—injection well pairs, relationships with a positive slope (i.e., increase in injection leads to increase in water level) are retained. As noted above, pumping rates for both extraction and injection are positive.

Next, the p-values associated with the data-model slope difference ($p_{\beta_{3_{cal_i}}}$) along with the slopes $\beta_{3_{cal_i}}$ are used to characterize the extent to which the modeled response matches the data response. For the case of a significant response in the data ($p_{\beta_{data}} < \alpha$), any model slopes with the opposite sign of the data slope are first excluded, leaving three options: model underestimates response ($p_{\beta_{3_{cal_i}}} < \alpha$; $\left|\beta_{3_{cal_i}}\right| < \left|\beta_{data}\right|$), model adequately matches response ($p_{\beta_{3_{cal_i}}} > \alpha$), and model overestimates response ($p_{\beta_{cal_i}} < \alpha$; $\left|\beta_{3_{cal_i}}\right| > \left|\beta_{data}\right|$). The absolute value notation is used because both extraction and injection wells are included in this analysis, and the slopes for injection well responses are negative. For the case of no significant response in the data ($p_{\beta_{data}} > \alpha$), there are two options: model matches lack of response ($p_{\beta_{3_{cal_i}}} > \alpha$), and model has response that is not observed ($p_{\beta_{3_{cal_i}}} < \alpha$).

3. Results and Discussion

3.1. Overall Summary

Well-pair classifications for all pumping wells and model calibrations are presented in Figure 4, and a full presentation of all Huber regressions and residuals can be found in Appendices A and B. $p_{\beta_{3_{cal_i}}}$ is the key parameter determining the share of well pairs classified broadly as matched vs. mismatched by the model, whereas $p_{\beta_{data}}$ further clarifies the type of match or mismatch. Overall, 84.5% of well pair relationships analyzed are matched by the model (upper quadrants in Figure 4), indicating that the model does not significantly differ from the data ($p_{\beta_{3_{cal_i}}} > \alpha$). That category is further divided into well pairs that do not have a significant connection evident from the data ($p_{\beta_{data}} > \alpha$, "Model Matches No Connection", 69%), and cases where a significant connection was identified ($p_{\beta_{data}} < \alpha$, "Model Matches Connection", 15.5%). The model differs significantly from the data ($p_{\beta_{3_{cal_i}}} < \alpha$) in 15.5% of well pairs analyzed (lower quadrants in Figure 4), with 7.5% of events being cases where the model contains a connection that is not evident in the data ($p_{\beta_{data}} > \alpha$, "Mismatch: Model Overly Connected"), and 7.5% being cases where the model overstates or understates a significant connection ($p_{\beta_{data}} < \alpha$, "Mismatch: Model Overstates or Understates Connection").

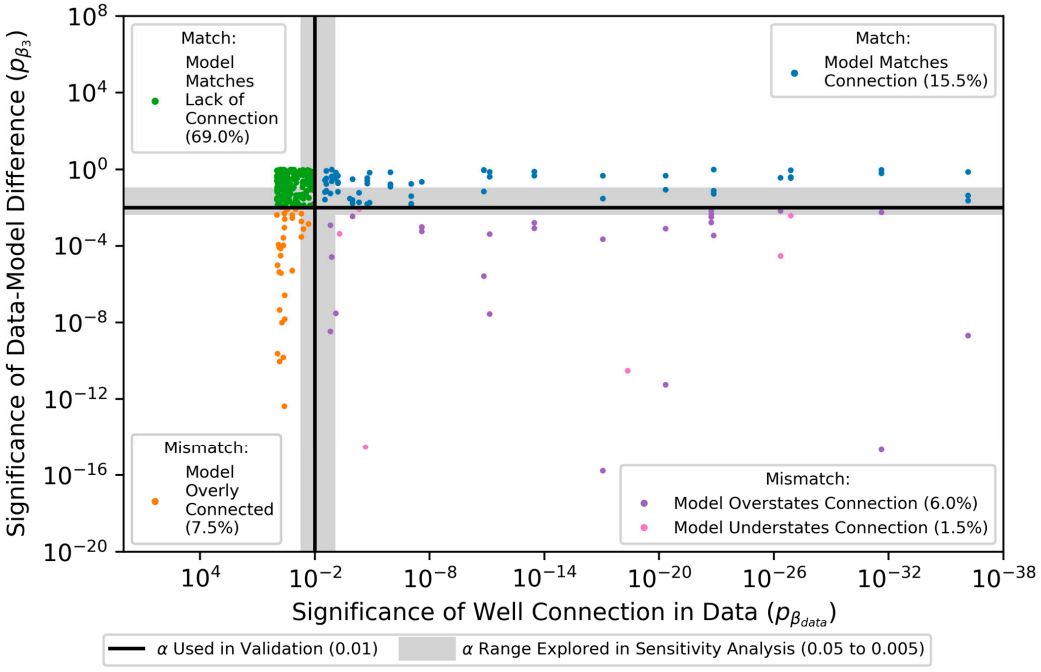

Figure 4. Well-pair classifications based on p-value criteria; significance levels denoted as black lines and gray shading. Note that the effect of adjusting α can be visually interpreted as shifts of the black

line within the gray shading. Well pairs with data-model difference *p*-values greater than 0.01 are classified as matches (upper quadrants, 84.5%), with the model reproducing significant connections on the upper right (blue, 15.5%) and no connection on the upper left (green, 69%). Well pairs with data-model difference *p*-values less than 0.01 are classified as mismatched (lower quadrants, 15.5%), with the model overstating (purple, 6%) or understating (pink, 1.5%) a significant connection on the lower right and the model simulating connections that are not observed in the data on the lower left (orange, 7.5%).

3.2. Sensitivity of Results to $p_{\beta_{3_{cal_i}}}$

As $p_{\beta_{3_{cal_i}}}$ is a key parameter in evaluating the model's ability to match the data, it is important to understand its influence on results. The sensitivity of validation results to the significance level (α) is presented in Table 3, and the effect α can be visually interpreted in Figure 4. α is used to characterize both the significance of well connections in the data (x-axis in Figure 4) and the data-model slope difference (y-axis in Figure 4). With higher α, more well connections are classified as significant in the data, and more data-model mismatches are identified. However, as the data-model slope difference ($p_{\beta_{3_{cal_i}}}$) is the key parameter determining the overall agreement of the model and data (y-axis in Figure 4), Table 3 focuses only on the effect of α on these classifications.

Table 3. Sensitivity of results to significance level ($\alpha = 0.01$ was used in validation results). Increasing α imposes a higher standard for data-model agreement and decreases the number of matches.

Data-Model Agreement ($p_{\beta_{3_{cal_i}}}$)	α = 0.005	α = 0.01	α = 0.05
Match	86.5%	84.5%	73%
Mismatch	13.5%	15.5%	27%

Increasing α imposes a higher standard for data-model agreement, and the share of well pairs classified as "Match" (Table 3) decreases from 86.5% ($\alpha = 0.005$), to 84.5% ($\alpha = 0.01$), to 73% ($\alpha = 0.05$). Similarly, the share of well pairs classified as "Mismatch" increases from 13.5% ($\alpha = 0.001$), to 15.5% ($\alpha = 0.01$), to 27% ($\alpha = 0.1$). A secondary effect of increasing α, of course, is to increase the number of well connections classified as significant in the data, but the primary influence of increasing α as it relates to classifying matched vs. mismatched connections is to increase the standard for data-model agreement. As discussed in Section 2, an appropriate value for α was selected based on an extensive review of individual events and well-pair classifications; in addition, the clusters evident in Figure 4 suggest that $\alpha = 0.01$ is appropriate for distinguishing between categories.

3.3. Validation Results for Well Pairs Used in the Development of Calibration Drawdown Targets

Figure 5 presents the same results as Figure 4, with the classifications divided by pumping well (CrEX-1 vs. remaining wells CrEX-2, -3 and CrIN-3, -4, and -5). For these results, it is important to understand some key details of the CM calibration. As discussed in Section 1, the calibration included drawdown targets for three wells (CrEX-1, CrEX-5, and CrIN-2) using opportunistic periods in which pumping is commenced while no other wells are active. Of these three drawdown target wells, only one (CrEX-1) is included in the validation, as CrEX-5 and CrIN-2 are excluded due to low sample sizes. The event itself is excluded from the validation to maintain separation between the calibration and validation datasets, but the well pairs involved in the drawdown target development are classified using the validation tool. Thus, CrEX-1 provides a unique example in which the validation performance can be compared to the calibration strategy (Figure 5). For the five monitoring wells (R-28, R-44 S1, R-44 S2, R-50 S1, and R-50 S2) that were used to develop CrEX-1 drawdown targets, the well-pair classifications are highlighted in cyan in Figure 5.

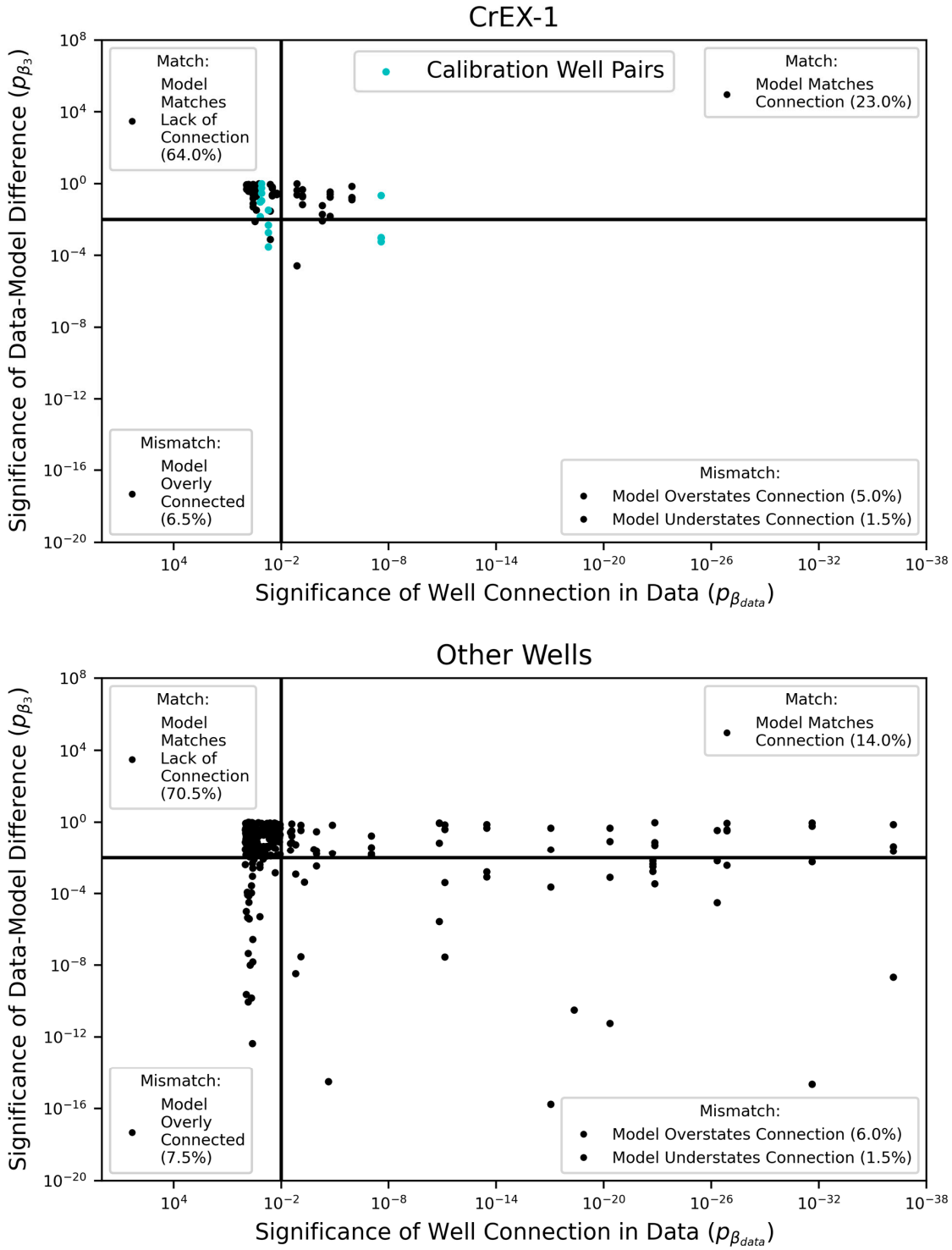

Figure 5. Well-pair classifications for CrEX-1 well pairs (**top**) and other wells (**bottom**); CrEX-2 and -3, CrIN-3, -4, and -5). CrEX-1 well pairs that were used to develop drawdown targets for the calibration are identified in cyan.

As CrEX-1 drawdown targets were explicitly included in the calibration, better validation results were achieved in CrEX-1 compared to the other wells. Indeed, CrEX-1 has significantly more instances of "Model Matches Connection" than the other wells (23% for CrEX-1 vs. 14% for the remaining wells), and more matches overall (87% vs. 84.5% for the remaining wells). Most of these matches, however, are in the opposite direction from the wells used in calibration drawdown targets (see Appendix C Figure A25), sug-

gesting that this success may be unrelated to the calibration drawdown targets (except for CrPZ-2, which is located nearby R-28). Figure 5 shows that most of the well pairs used to develop CrEX-1 drawdown targets have $p_{\beta_{data}} > 0.01$ (left quadrants), and just four points have $p_{\beta_{data}} < 0.01$ (right quadrants). These four are associated with the significant connection between CrEX-1 and R-50 S2: calibration 3 matched this connection, whereas the other calibrations overstated the response (by a factor of six). The remainder of the well pairs were classified as not connected in the data, highlighting differences between the pumping conditions under which the validation was conducted (many wells pumping at once) vs. those of the drawdown target periods (only a single well pumping).

To understand these differences further, a map-based representation of connection classifications for monitoring wells used in the development of calibration drawdown targets is presented in Figure 6. Figure 6 summarizes the results across calibrations based on the most common outcome for the monitoring well to emphasize the broad behavior of the model across calibrations with respect to the data; however, each calibration is classified independently (see Appendix A Figure A1). Five monitoring well screens were involved in developing CrEX-1 drawdown targets in the calibration: R-28, R-44 S1 and S2, and R-50 S1 and S2. In terms of the data-based classifications, except for R-50 S2 (discussed above), these wells were either classified as not connected (R-28, R-44 S2, R-50 S1) or as demonstrating connections to injection (R-44 S1). Of the wells that were classified as not connected in the data, simulated responses either matched the lack of connection (R-44 S2; R-50 S1) or overstated the connection (R-28).

Considering the locations of R-50 and R-44 (Figure 6), it is clear these monitoring wells are much closer to the array of injection wells (CrIN-3, -4, and -5) than CrEX-1, and R-44 S1 and R-50 S1 are both connected to CrINs 3–5 in both the data and the model calibrations. Whereas CrEX-1's connection to these wells is measurable in the rare circumstances during which the CrINs are not in operation, this connection is generally dominated or hidden by the effects of the CrINs when they begin injecting. The case of R-28 is similar but related to CrEX-3 rather than the CrINs. Figure 7 shows that CrEX-3 is located directly between CrEX-1 and R-28 and thus interferes with the CrEX-1, R-28 connection in an operational setting. The CrEX-3, R-28 connection, then, is classified as significant and dominates the connection between CrEX-1 and R-28.

Stepping back, the cases of R-28 (model overly connected) and R-50 S2 (overstate significant connection) are especially interesting and suggest that the choice of calibration drawdown targets has unduly influenced the model due to the unusual nature of the conditions in which the drawdown target occurs. Such comparisons between the calibration and validation results reflect differences in purpose and methodology: calibration takes advantage of unique operational settings to derive hydraulic flow properties, whereas the validation tool identifies the dominant well connections and the ability of the model to reproduce them during normal pumping operations. Whereas calibration seeks a value of hydraulic conductivity between CrEX-1 and R-28 to parameterize that region of the subsurface explicitly in the numerical model, the validation focuses instead on the CrEX-3, R-28 connection that dominates during complex operational conditions. This connection, incidentally, is overstated in the model compared to the data, likely due to the CrEX-1, R-28 drawdown target. Similarly, R-50 S2's relationship with CrEX-1 is generally muted during normal operating conditions due to the injection of the CrINs but calibrating the model to explicitly represent this connection has the effect of overemphasizing its importance. Overall, the validation tool provides a unique opportunity to explore the influence of calibration targets on the calibrated flow field during normal operational conditions and suggests that the choice of a particular set of well pairs to target in the calibration (often based on data availability) creates flow pathways that may be inconsequential or irrelevant for the conditions during which the model is making predictions. Although calibration is often portrayed as a search for intrinsic system properties, such parameters inevitably contain artifacts of the conditions during which the observations used to derive them were made. This validation method allows the modeler to directly interrogate these calibration assumptions and improve the model in an iterative manner.

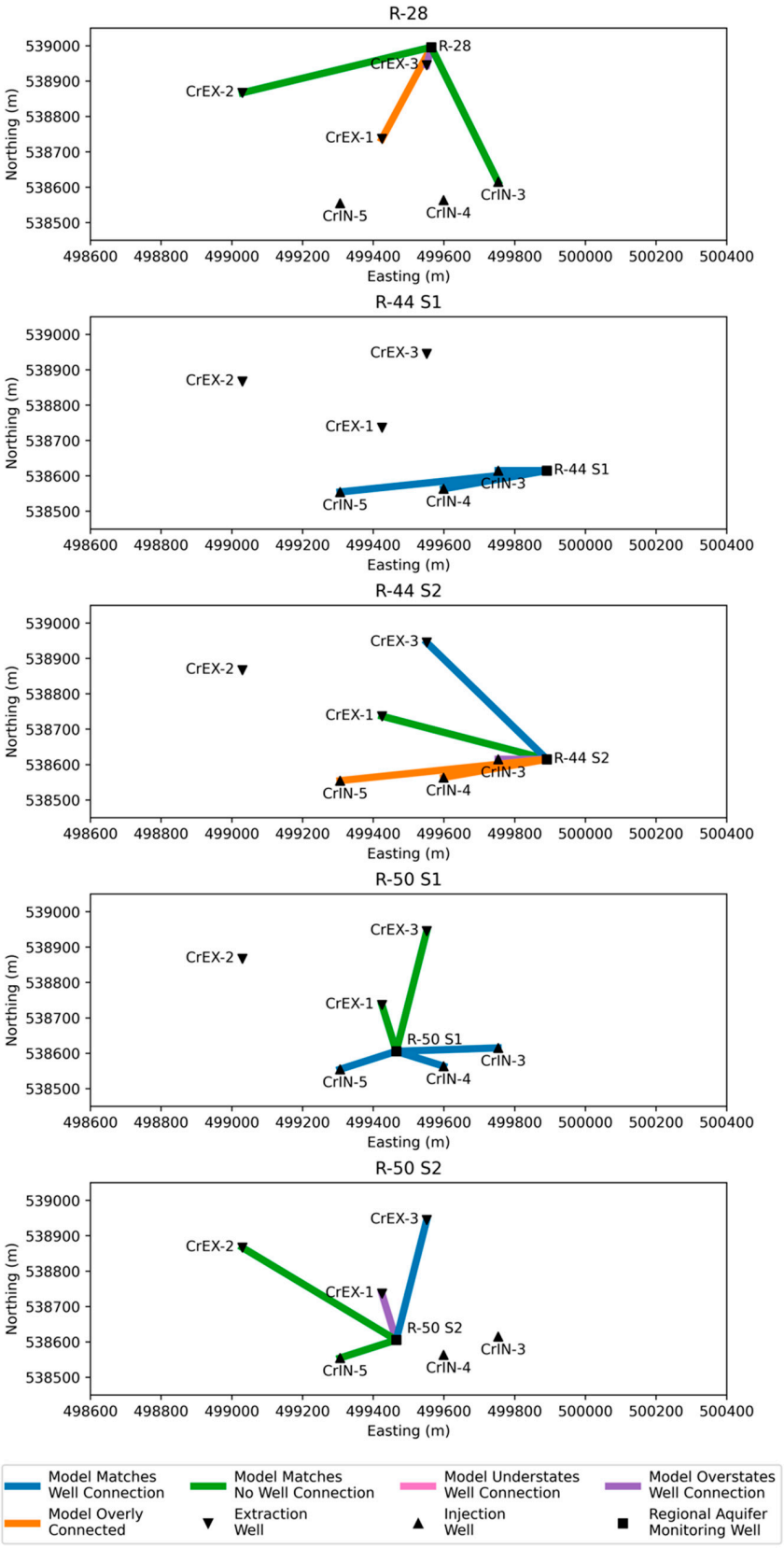

Figure 6. Map-based representation of connection classifications for monitoring wells used in the development of drawdown targets during model calibration. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

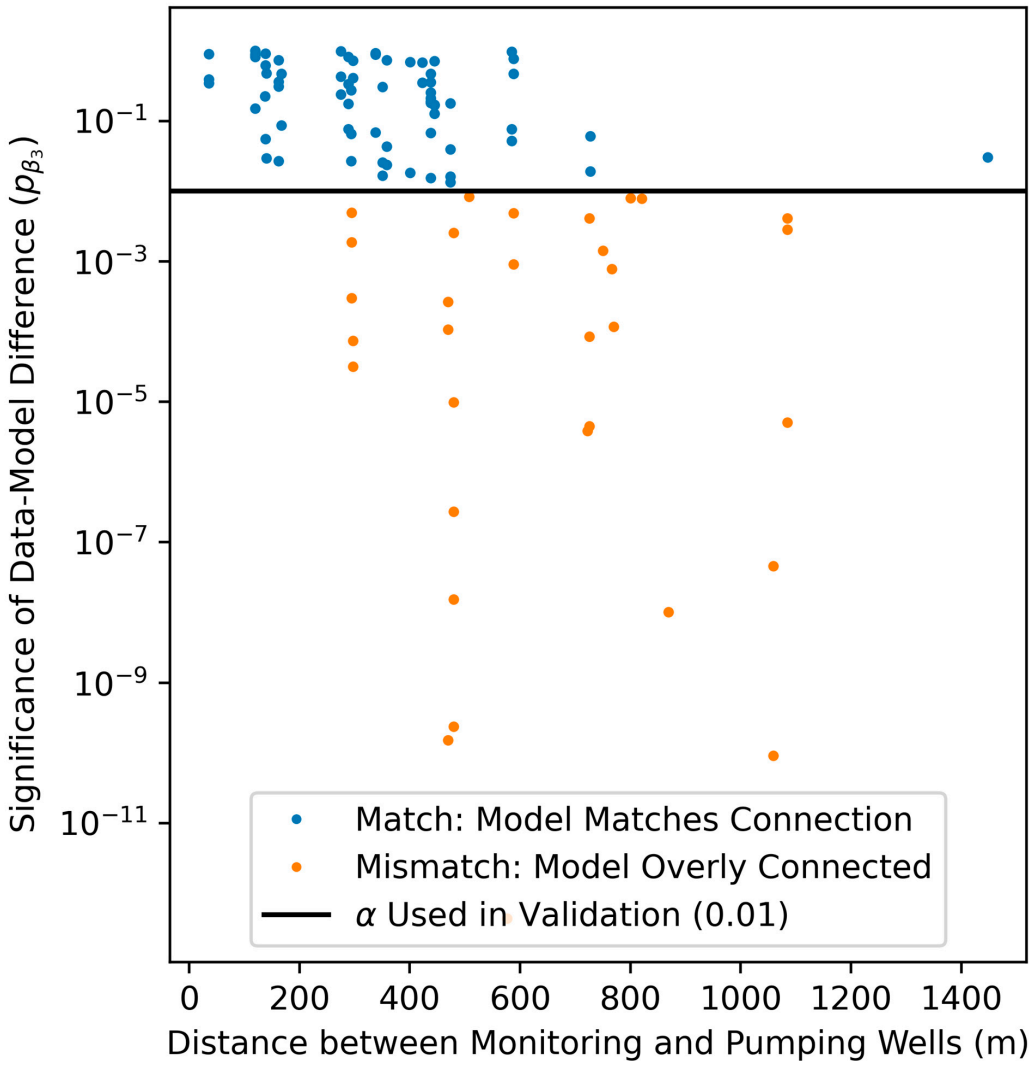

Figure 7. Well-pair classifications based on *p*-value criteria against distance between monitoring and pumping wells. Instances of "Model Overly Connected" tend to occur at distances of 250–1750 m, whereas instances of "Model Matches Connection" occur at shorter distances (<750 m).

3.4. Spatial Trends in Well-Pair Classifications

To examine the spatial distribution of model performance, the "Model Overly Connected" and "Model Matches Well Connection" categories are compared in Figure 7 as a function of distance between the well pairs. Figure 8 shows the same categories in a map-based representation against an estimate of the Cr plume boundary drawn in 2022 for additional spatial context. Figure 7 shows that the model adequately matches connections over shorter distances (<800 m), and instances of over-connectedness occur over greater spatial scales (300–1200 m). As discussed in Section 1, groundwater models are required to simulate subsurface flow as well as the lack of subsurface flow. In this context, Figure 7 suggests that the over-representation of flow generally occurs over greater distances. Such a problem is likely widespread among groundwater models and can readily be resolved through the introduction of "null" drawdown target observations in the calibration.

Figure 8 shows that the model tends to match well connections along the southern boundary of the plume, whereas the instances of over-connectedness predominantly occur along the northeastern plume edge. This region was identified through a capture zone analysis [36] as exhibiting a more uncertain flow regime than the southern and southeastern plume boundaries and would benefit from additional observations and parameter characterization in the model.

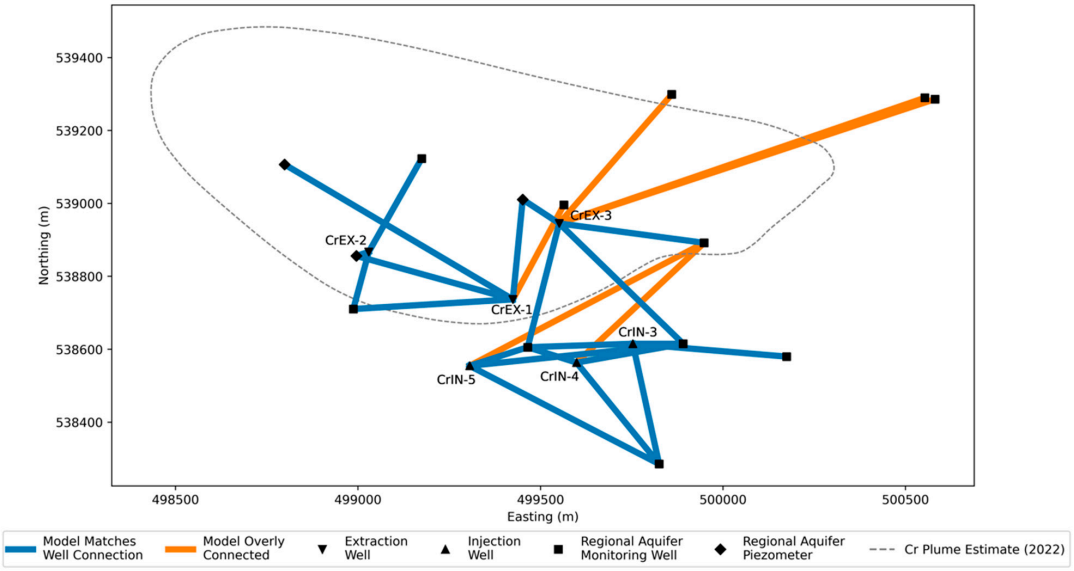

Figure 8. Map-based representation of response classifications for CrEX-1 well pairs. Similar to Figure 6, line colors summarize results across calibrations based on the most common outcome for the monitoring well as a way to understand the broad behavior of the model with respect to the data. The general direction of flow is southeast (from the top left of the figure to the bottom right).

With respect to the general direction of groundwater flow (southeast), over-connectedness tends to occur orthogonally (northeast), whereas matches tend to occur parallel to groundwater flow (southeast). Such a trend could indicate that the model is over-relying on advective transport as opposed to dispersion to match concentration targets in the transport model (i.e., perhaps hydraulic conductivity is set too high, whereas dispersivity is set too low, though other physical parameters and flow field nuances could produce this same result). Over-connectedness is associated primarily with CrEX-1, CrEX-3, CrIN-4, and CrIN-5. While the calibration has, in principle, optimized the dispersion parameters, it is possible for the calibration algorithm to settle on a large hydraulic conductivity to achieve concentration targets due to the lack of "null" drawdown targets. Broadly, these spatial comparisons demonstrate the utility of the validation tool in identifying regions that would benefit from additional observations, as well as areas for model improvement. More than a simple statistical score, the tool provides spatial distinctions (northeastern vs. southeastern plume edge), identifies wells whose influence is over-represented in the model (CrEX-1, CrEX-3, CrIN-4, CrIN-5), and distinguishes the spatial scales over which the model matches subsurface flow (<800 m) vs. the spatial scale associated with over-connectedness (300–1200 m).

4. Conclusions

In this study, we presented a groundwater model validation tool using robust multiple regression of pumping events and water level responses for wells that compose a contaminant pump-and-treat system. Four deterministic calibrations of the FEHM-based Chromium Model (CM) were used as a validation test case and successfully matched connections between wells in 84.5% of the well pairs analyzed. That the model was calibrated against only six of these well pairs (4% of total well pairs analyzed) demonstrates the strength of the model in representing monitoring well responses that were not explicitly included in the calibration with drawdown targets.

Overall, the well with calibration drawdown targets (CrEX-1) performed better in validation than the remaining wells (23% "Model Matches Connection" for CrEX-1 vs. 14% for the remaining wells), suggesting that the effect of the drawdown targets was to improve the model. However, the validation tool identified two well pairs (CrEX-1, R-50 S2 and CrEX-1, R-28) whose importance was overemphasized during calibration, and suggested an

alternate set of monitoring wells that are more relevant for operational conditions (CrPZ-1, CrPZ-2, CrPZ-4, R-61 S1, and R-61 S2) and may serve as a more appropriate choice for subsequent iterations of calibration. The tool demonstrated that the model adequately matches the flow over short spatial scales (<800 m) and over-represents the flow over greater distances (300–1200 m). The issue of over-connectedness could readily be improved through the inclusion of "null" drawdown targets in the calibration for monitoring wells that are far from the pumping wells.

More broadly, the results of the model validation affirm the importance of applying independent methods to evaluate the results of model calibration, interrogate model assumptions (both stated and unstated), and verify that the choice of calibration targets does not unduly influence the model in unintended ways or contain artifacts of the conditions during which observations were selected. It is important not only to use an independent set of data to validate the models but also to use a diverse set of methods that differ fundamentally from the methods used to develop and calibrate the model. While multiple linear regression using a Huber loss function was selected for the current study based on exploratory data analysis and the need to reduce the influence of outliers, numerous modern machine learning methods could be similarly applied as a tool for independent model validation. Indeed, the field of groundwater contaminant hydrogeology has only begun to take advantage of the latest deep-learning models for numerical model validation, and future work should consider a wide array of tools for this purpose. Whereas the expense required to obtain deep subsurface observations is unlikely to decrease, our ability to forage for hidden information in scarce datasets can only improve by bringing the latest methods to bear on traditional groundwater contaminant problems.

The validation presented here is modeling platform-agnostic, makes few assumptions about the underlying system behavior, and could be broadly applied to any model that follows a predictable set of rules about its response to stresses, e.g., pumping leads to water level drawdown and ceasing pumping leads to water level recovery. The approach makes a set of pragmatic assumptions about the nature of water level responses over the timescale of the events considered. Specifically, by looking at only the change in water level over a timespan relevant to aquifer recovery, any long-term water level trends (e.g., due to recharge) can be neglected, and the non-linear processes of drawdown and recovery can be effectively reduced to a set of linear equations. This approach has the benefit of avoiding assumptions about the underlying physics describing groundwater flow (i.e., confined, leaky confined, unconfined) and the nature and causes of long-term water level fluctuations (e.g., due to recharge). These assumptions and the overall approach were inspired by monitoring well water levels that rapidly respond to pumping rate changes and approach steady-state drawdown and recovery relatively quickly (<7 days). Although the method could readily be applied to a system in which water levels do not respond as quickly, greater care would need to be taken to ensure that water level responses (Δh) for the data and model calibrations are computed over comparable timescales. For example, model output may be required at the same frequency (e.g., hourly) as water level observations. In addition, the minimum event length would likely need to be increased from two days to give the well sufficient time to proceed past the rapid initial response phase, when water levels may not be as comparable between the data and model.

Unlike many statistical scores (e.g., RMSE, Nash–Sutcliffe Efficiency, Coefficient of Determination), the validation tool provides nuanced model performance information that can directly be used in model refinement, such as detecting mechanisms that are inadequately represented by the model and delineating the distances over which the flow is adequately characterized vs. overly represented. Validation is, of course, part of "an iterative calibration-modeling-monitoring-evaluation-refinement cycle" [3]. In that spirit, the detailed information obtained by the validation tool regarding the Chromium Model is directly being used in ongoing recalibration and model improvement. The findings of this study constitute an especially rigorous review of the model calibration, demonstrate a culture of transparency, and reflect the importance of the model's predictions and its

influence on environmental health and policy. Indeed, one of the most valuable forms of knowledge is to understand its limits, and more rigorous validation can only serve to improve our models and their predictions.

Author Contributions: Conceptualization, M.R.; methodology, M.R., L.G.-G. and T.S.; software, M.R. and L.G.-G.; validation, M.R., L.G.-G. and T.S.; formal analysis, M.R.; investigation, M.R.; resources, L.M.F. and T.S.; writing—original draft preparation, M.R.; writing—review and editing, M.R., L.M.F., L.G.-G. and T.S.; visualization, M.R., L.M.F., L.G.-G. and T.S.; supervision, L.M.F. and T.S.; project administration, L.M.F. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based upon work supported by the Environmental Management Los Alamos Field Office of the Department of Energy.

Data Availability Statement: All data supporting the results of this study can be found in the Intellus New Mexico database: https://intellusnm.com/ (accessed on 8 November 2022).

Acknowledgments: The authors would like to thank Amy Jordan, Christopher Peck, Dylan Boyle, Daniel Birdsell, Gregg Occhiogrosso, and other members of the Chromium team at Neptune and Company, Inc., who provided valuable insights into the development of this method and subjected their work to the heightened scrutiny required for a validation exercise.

Conflicts of Interest: Michael Rush, Leslie Gains-Germain, Lauren M. Foster, and Tom Stockton are employed by Neptune and Company, Inc. This material is based upon work supported by the Environmental Management Los Alamos Field Office of the Department of Energy. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Exploratory Analysis and Analysis of Residuals

Prior to statistical modeling, the transformed data are visualized to assess general patterns. First, the relationship between the water level response and the pumping rate change is plotted in Figures A1–A6. These plots show that the relationship between the water level response and the pumping/injection rate change can generally be described by a linear relationship within each category, with no clear evidence for the need for a quadratic model or some other parametric or non-linear form. Though the sparsity of data in the middle of the scatterplot (with little to no pumping or injection) makes it difficult to evaluate the assumption of linearity, the estimated slopes adequately capture the relationships observed in the plots. The sparsity of data in the middle of the scatterplot is due to a lack of small pumping rate change events: when wells turn on or off, they typically pump at rates within a narrow range (e.g., 50–80 gpm), leading to two main clusters of pumping rate changes and water level responses constraining both ends of the linear model, without many smaller changes helping to constrain the middle. Specifically, most pumping changes are generally clustered between one and two standard deviations from zero, whereas water level responses tend to remain within one standard deviation from zero.

Next, the ratio of the water level response Δh to the pumping rate change Δq is visualized in a categorical fashion (Figures A7–A12) to compare the mean and variance of the responses in the data and model (calibrations 1–4). An average response ratio close to 0 indicates no water level response on average for that well pair and category. Since Δq is positive for both extraction and injection, a negative ratio indicates the expected response for extraction wells (increase/decrease in extraction leads to drawdown/recovery), whereas a positive ratio indicates the expected response for injection wells (increase/decrease in injection leads to increase/decrease in water level). Figures A7–A12 show that the average response ratio of the data often matches the average response ratio from the model calibrations. For example, in Figure A7, the ratios are all centered at 0 for each category at R-43 S1, though there are several outlier responses from the model calibrations. Other observation wells show substantial differences across data or calibration. For example, in Figure A7, the average ratio at R-44 S2 is positive for the data and calibrations 1, 2, and 3, indicating a response to injection, whereas calibration 4, however, has a negative average

ratio, indicating a response to extraction. Additionally, the variance in $\Delta h/\Delta q$ ratios can differ substantially between data and model calibrations. At R-33 S1 in Figure A7, the variance is substantially larger in the data. In contrast, the variance at R-11 and R-35b is larger in the model calibrations compared to the data. Typically, model calibrations have more outliers than the data. The data distribution appears to be more normal, with occasional outlying observations at a frequency expected from random sampling. The distribution of model response ratios is less consistent, with more frequent outlying observations, such as the three relatively large and negative ratios for R-28 in A7.

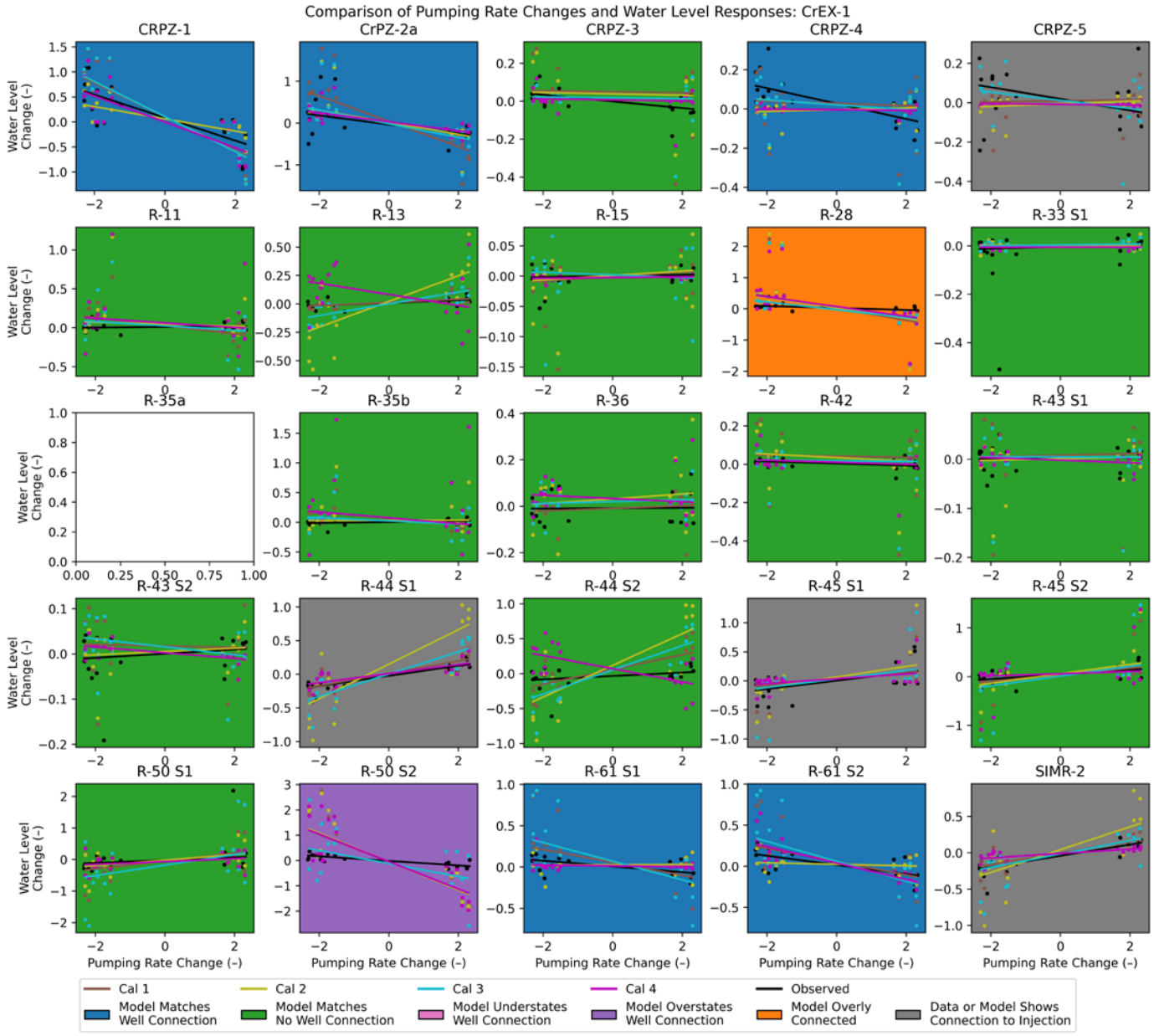

Figure A1. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrEX-1 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. R-35a is excluded from the CrEX-1 analysis due to a low sample size (<15 events). Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

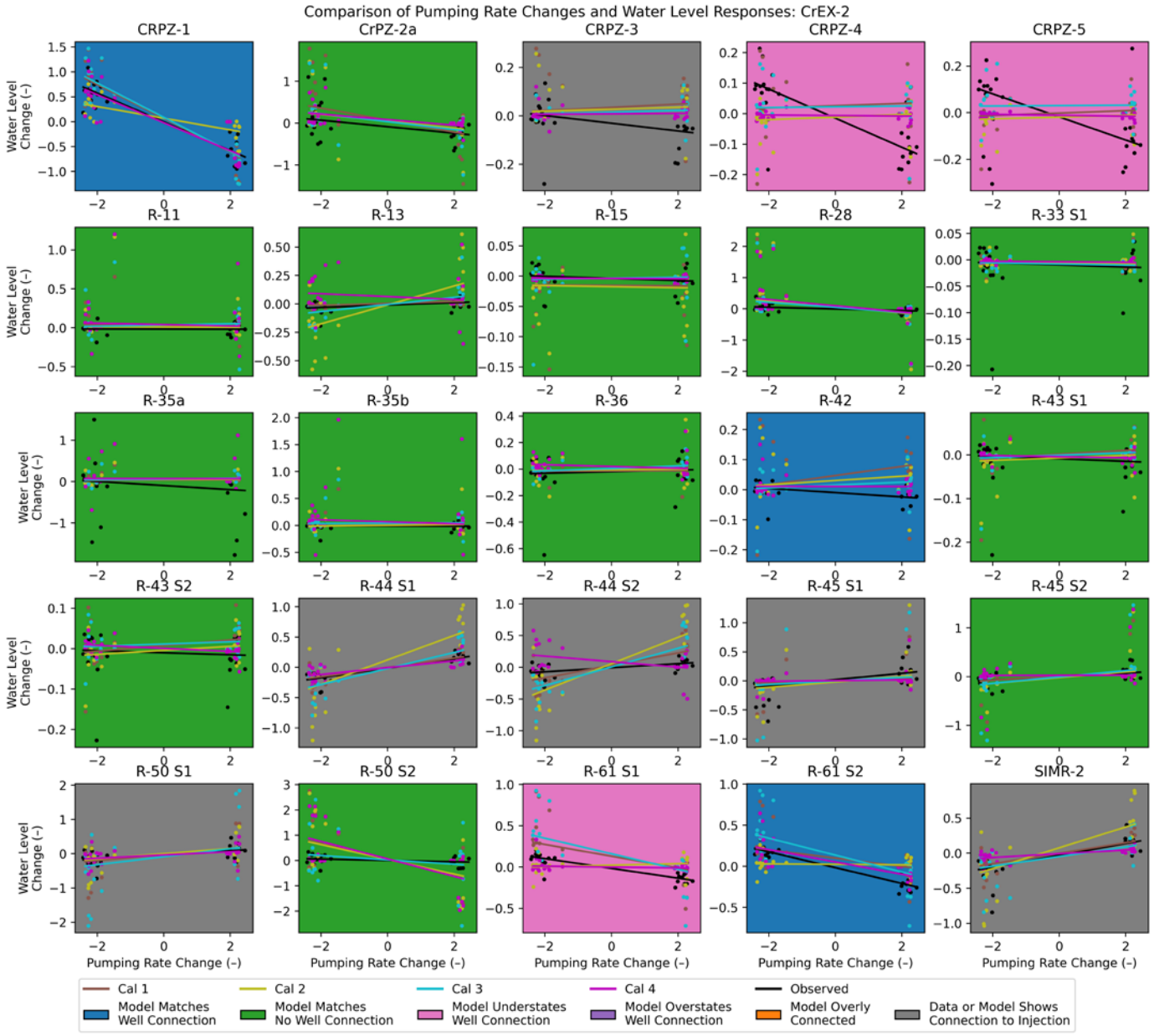

Figure A2. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrEX-2 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

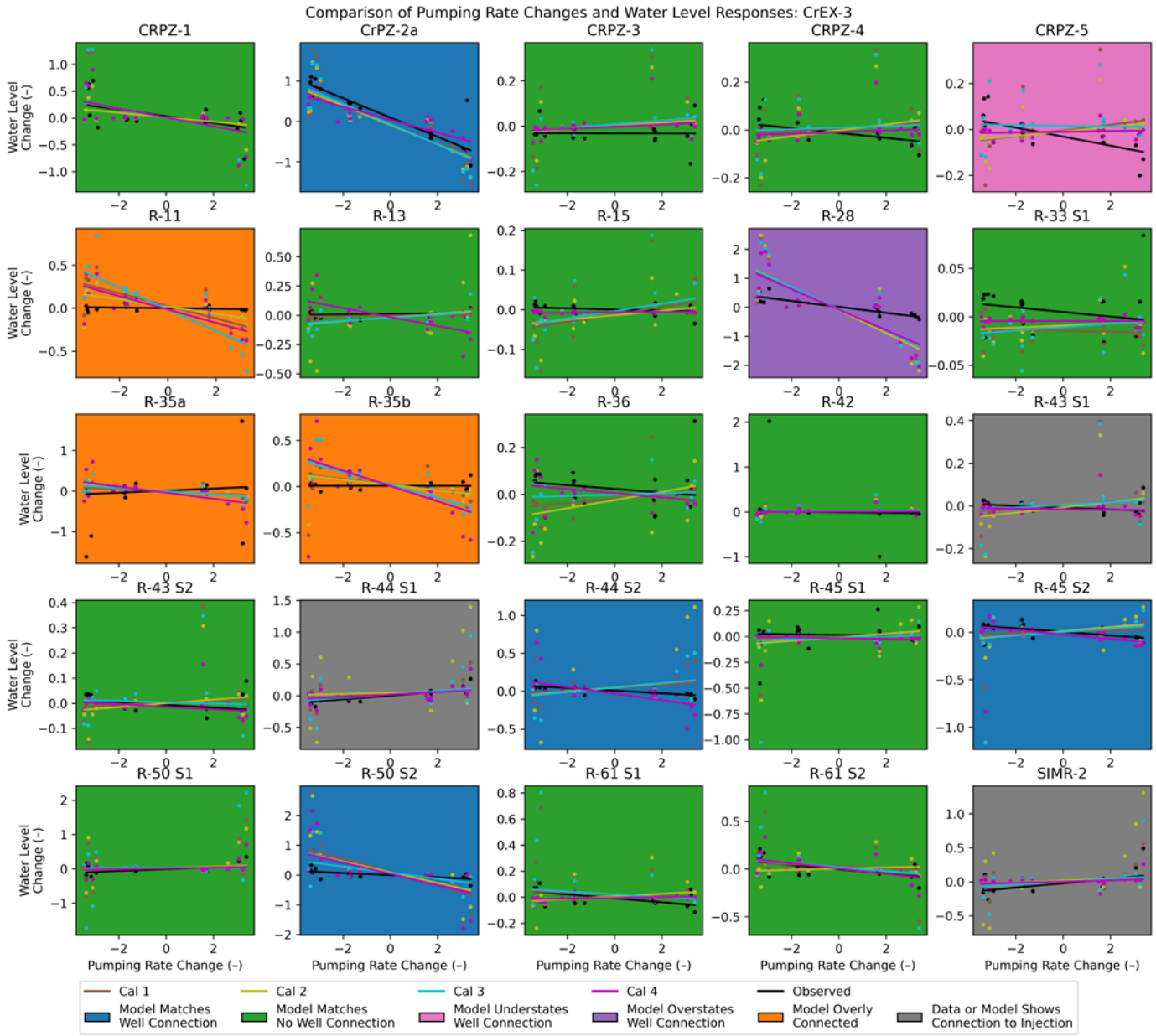

Figure A3. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrEX-3 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

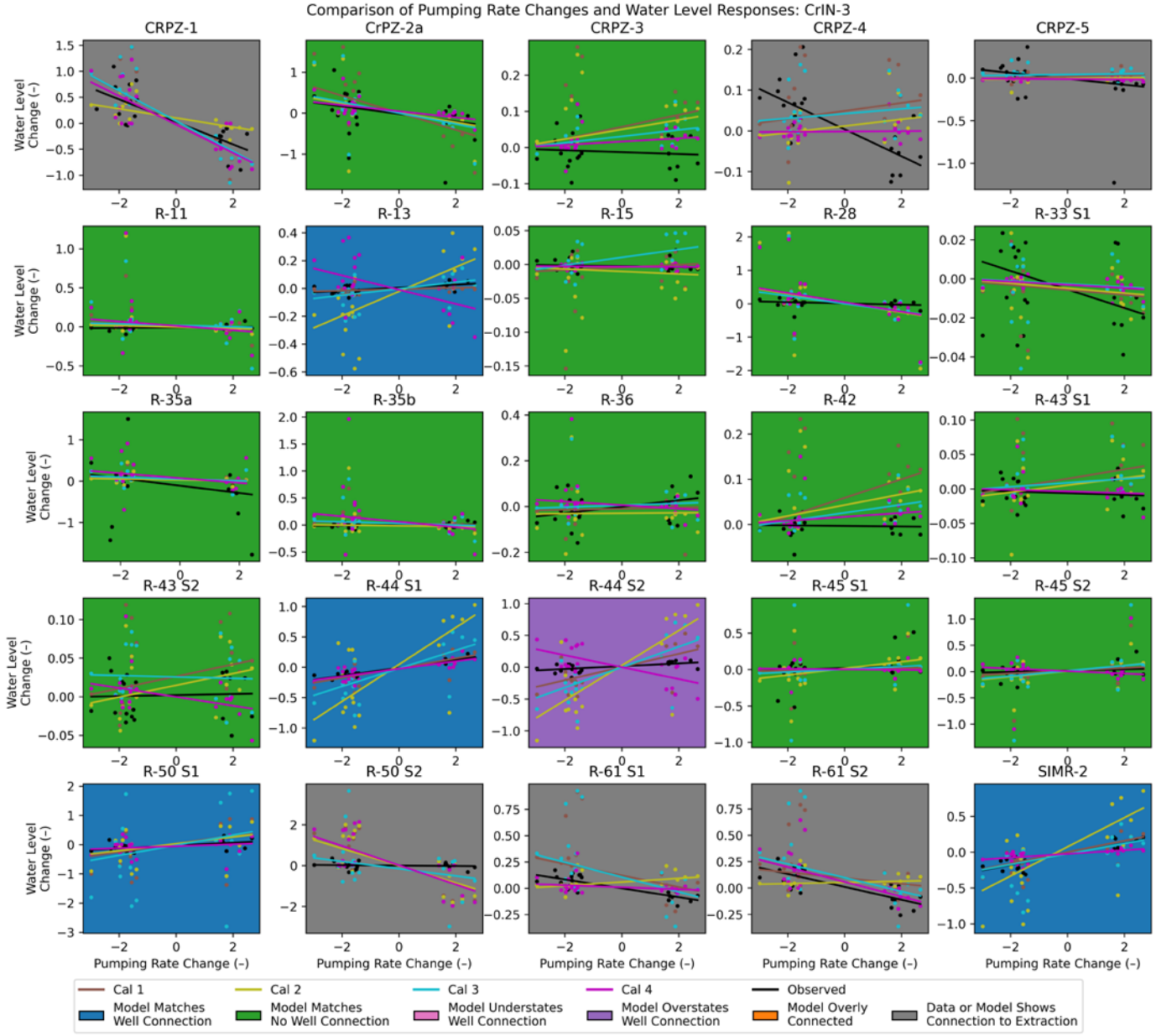

Figure A4. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrIN-3 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

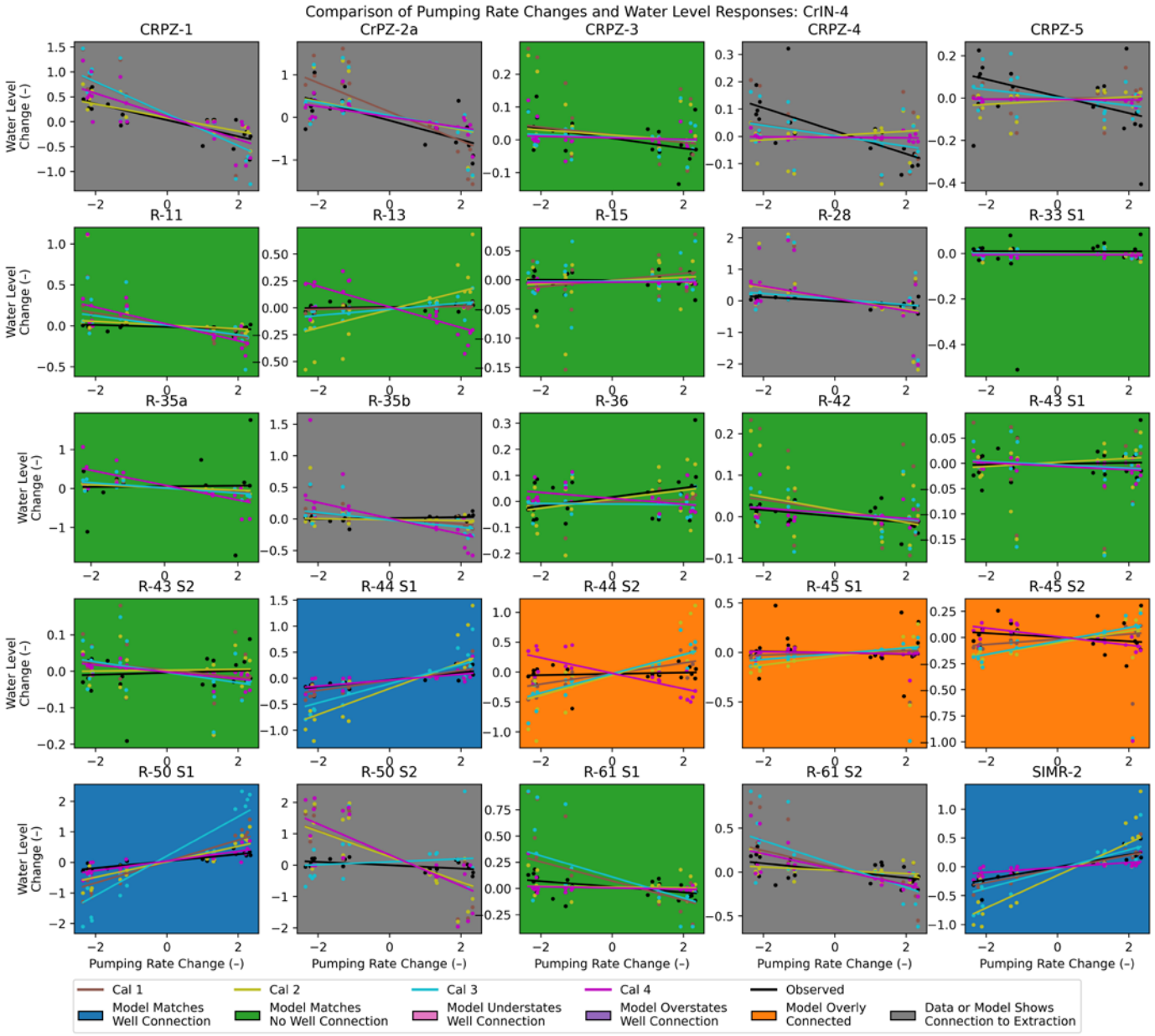

Figure A5. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrIN-4 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

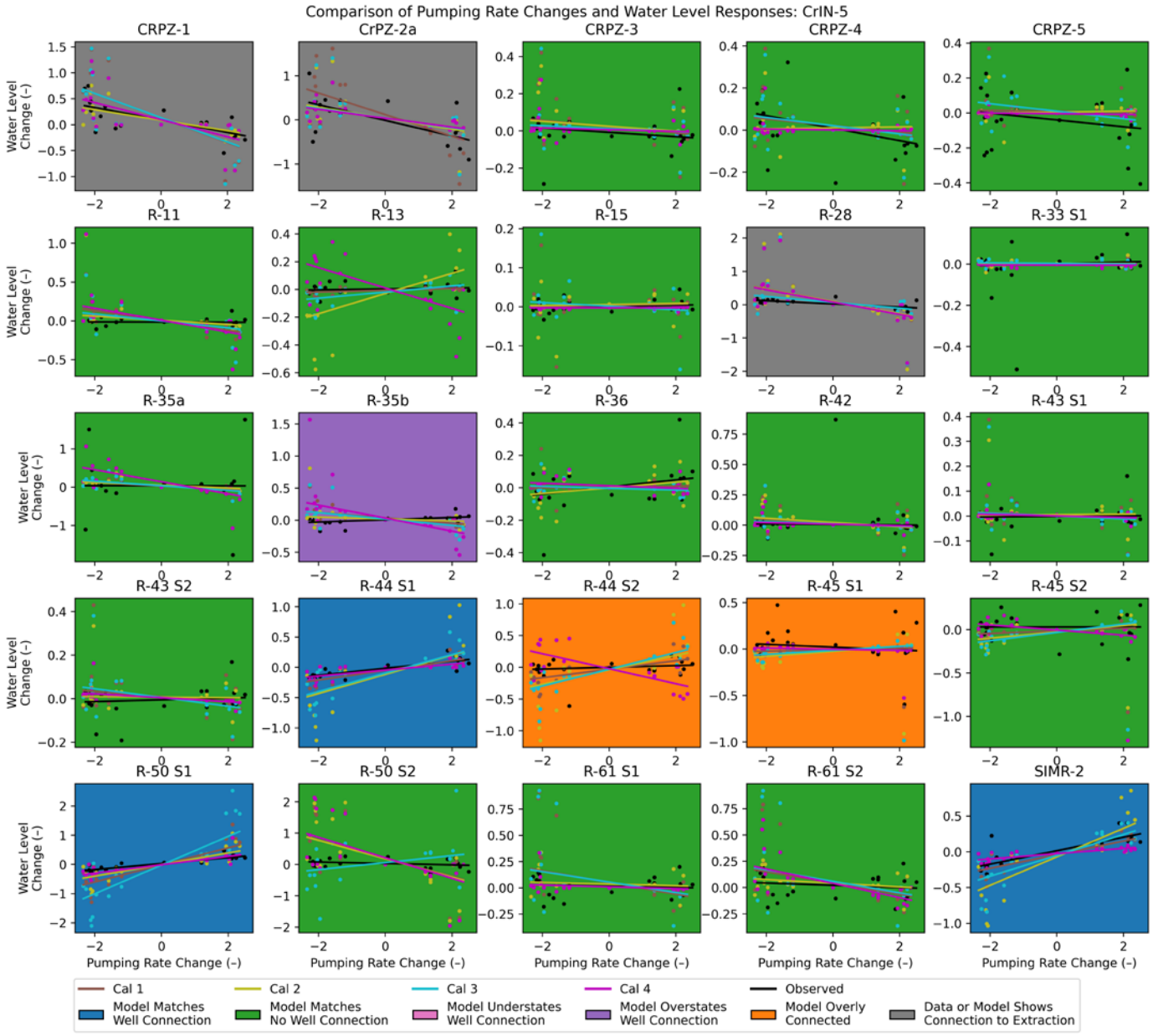

Figure A6. Scatter plots and Huber regressions of pumping rate changes and monitoring well responses for CrIN-5 well pairs. Plot colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data. Each point on each panel represents Δq and Δh for a single event for the given well pair. Δq and Δh are unitless because they are normalized by their respective standard deviations.

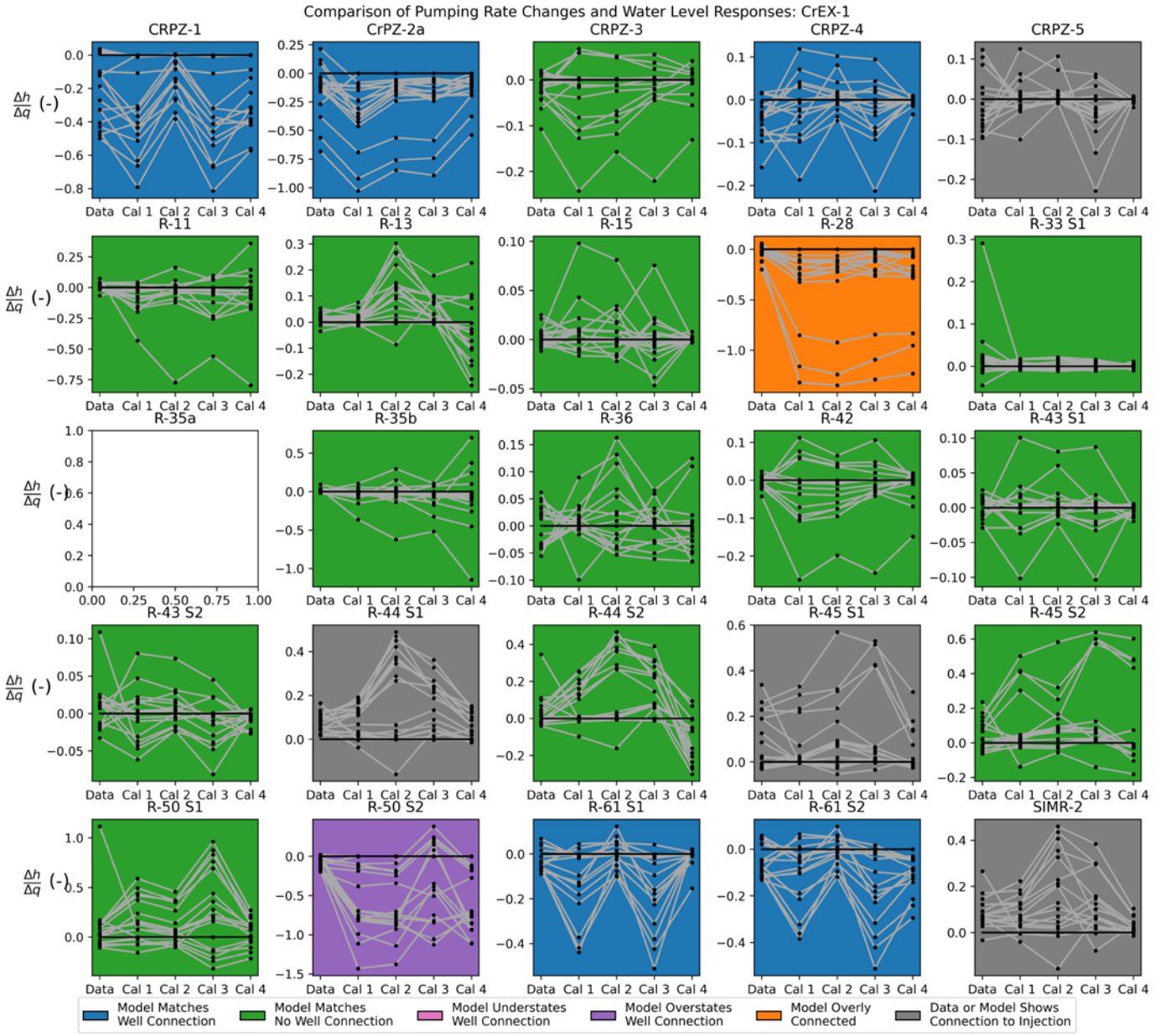

Figure A7. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrEX-1 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

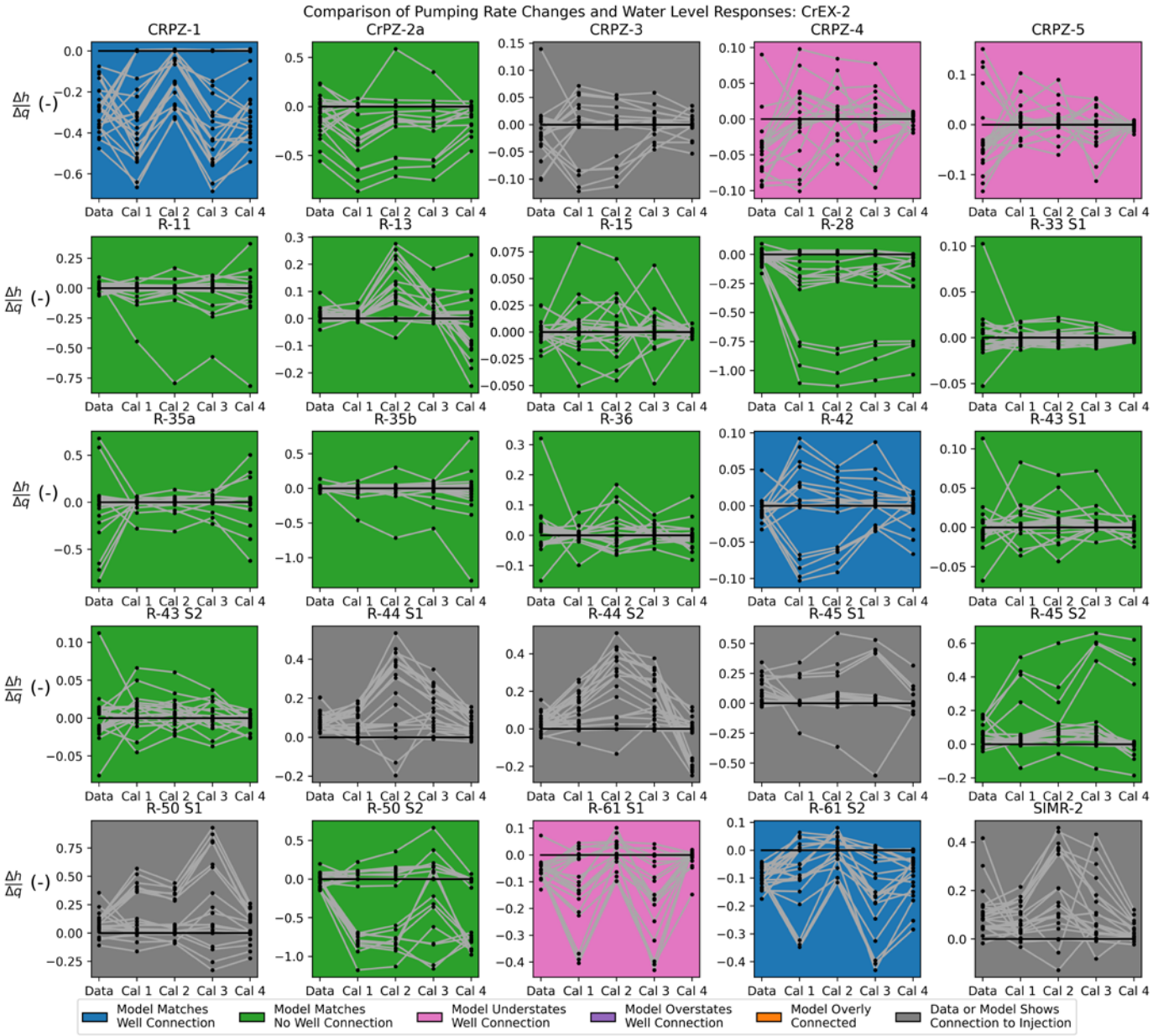

Figure A8. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrEX-2 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

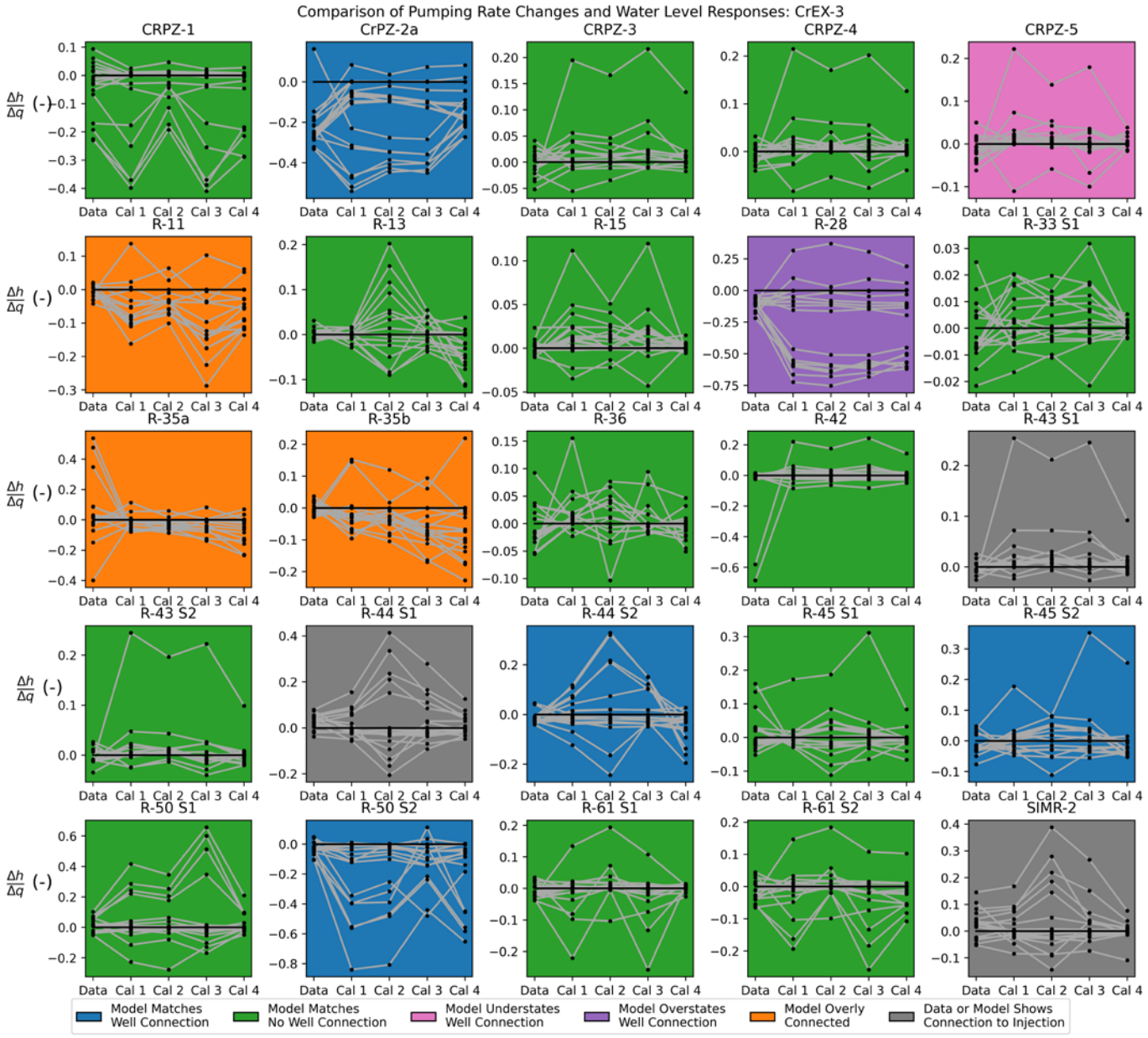

Figure A9. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrEX-3 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

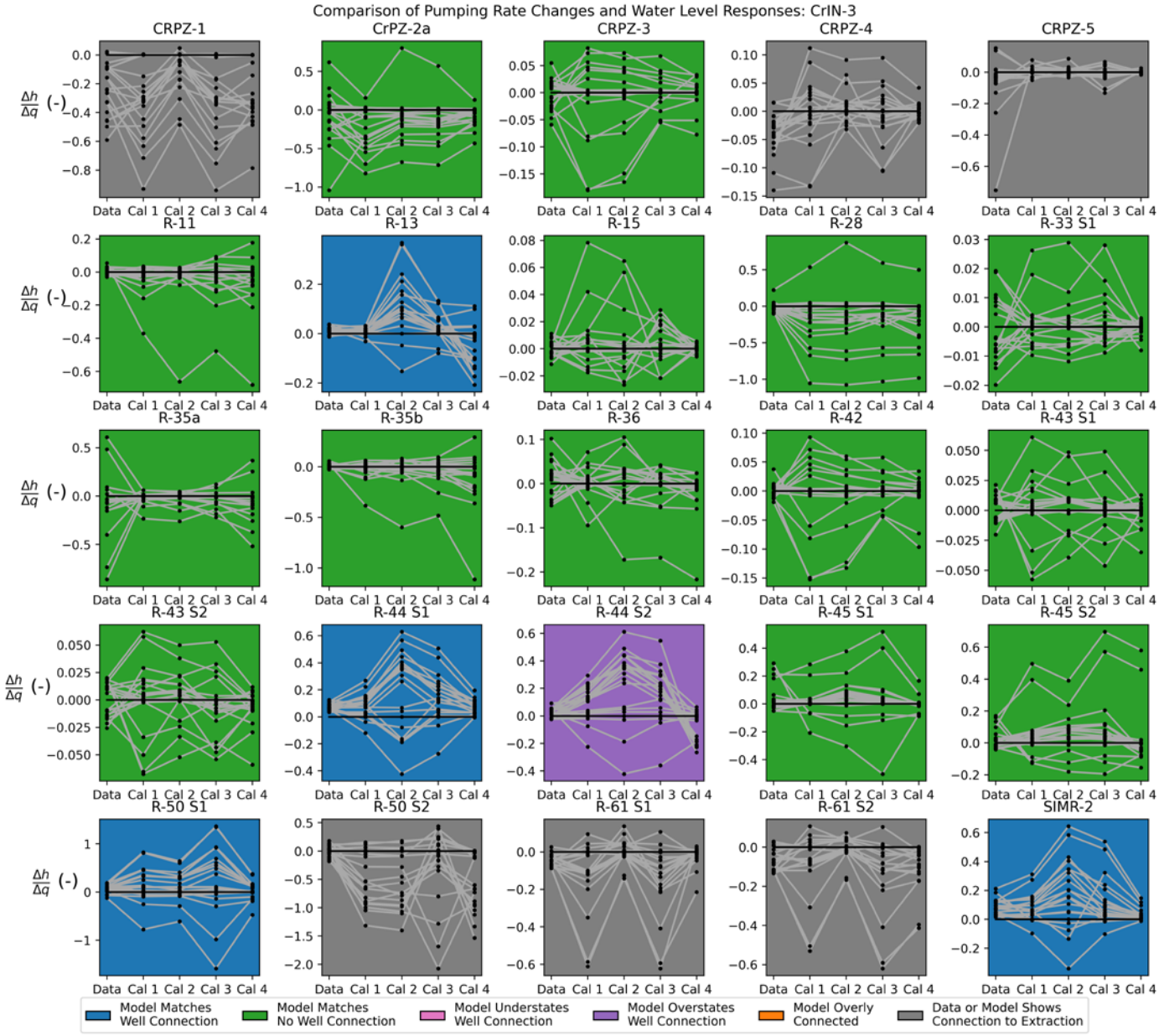

Figure A10. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrIN-3 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

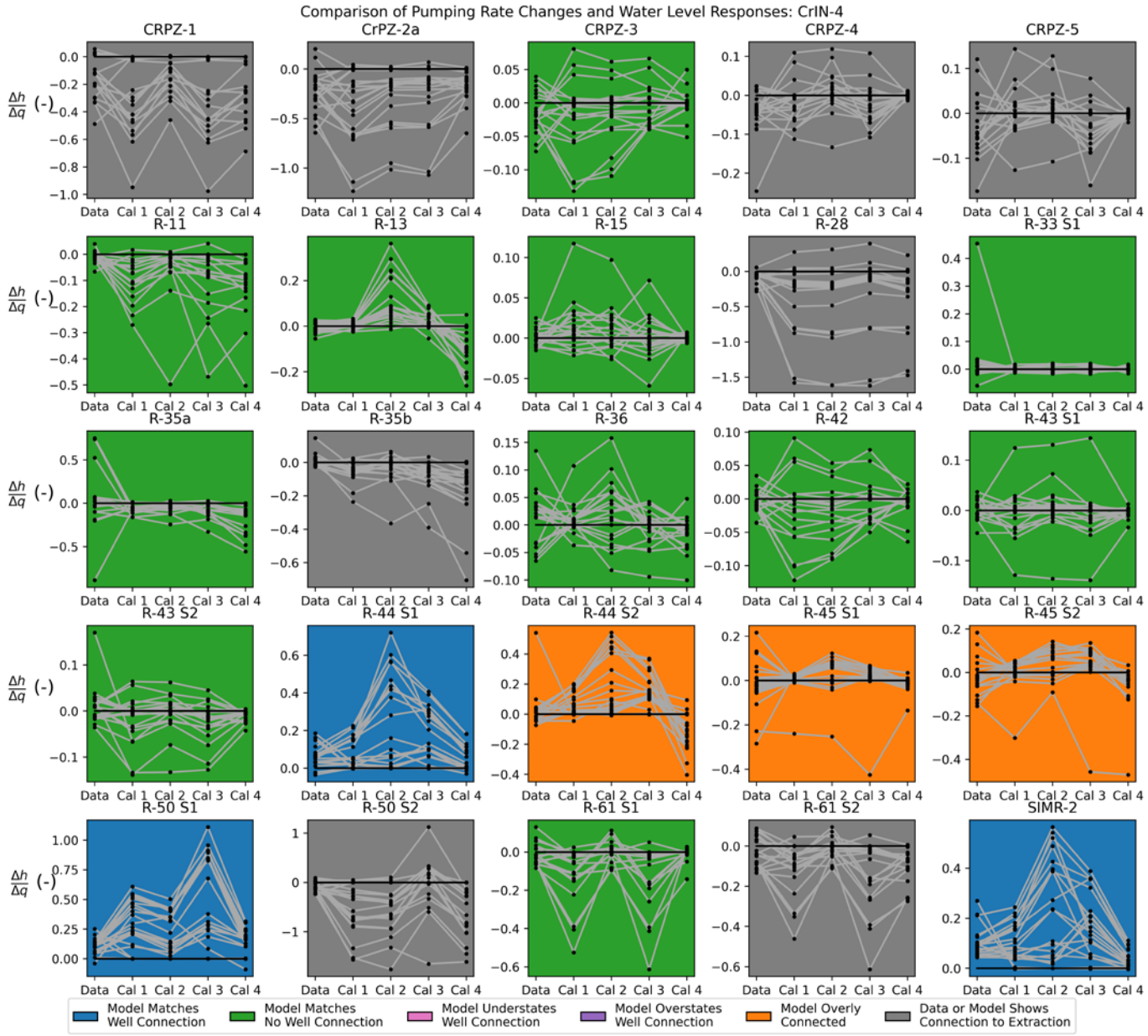

Figure A11. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrIN-4 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

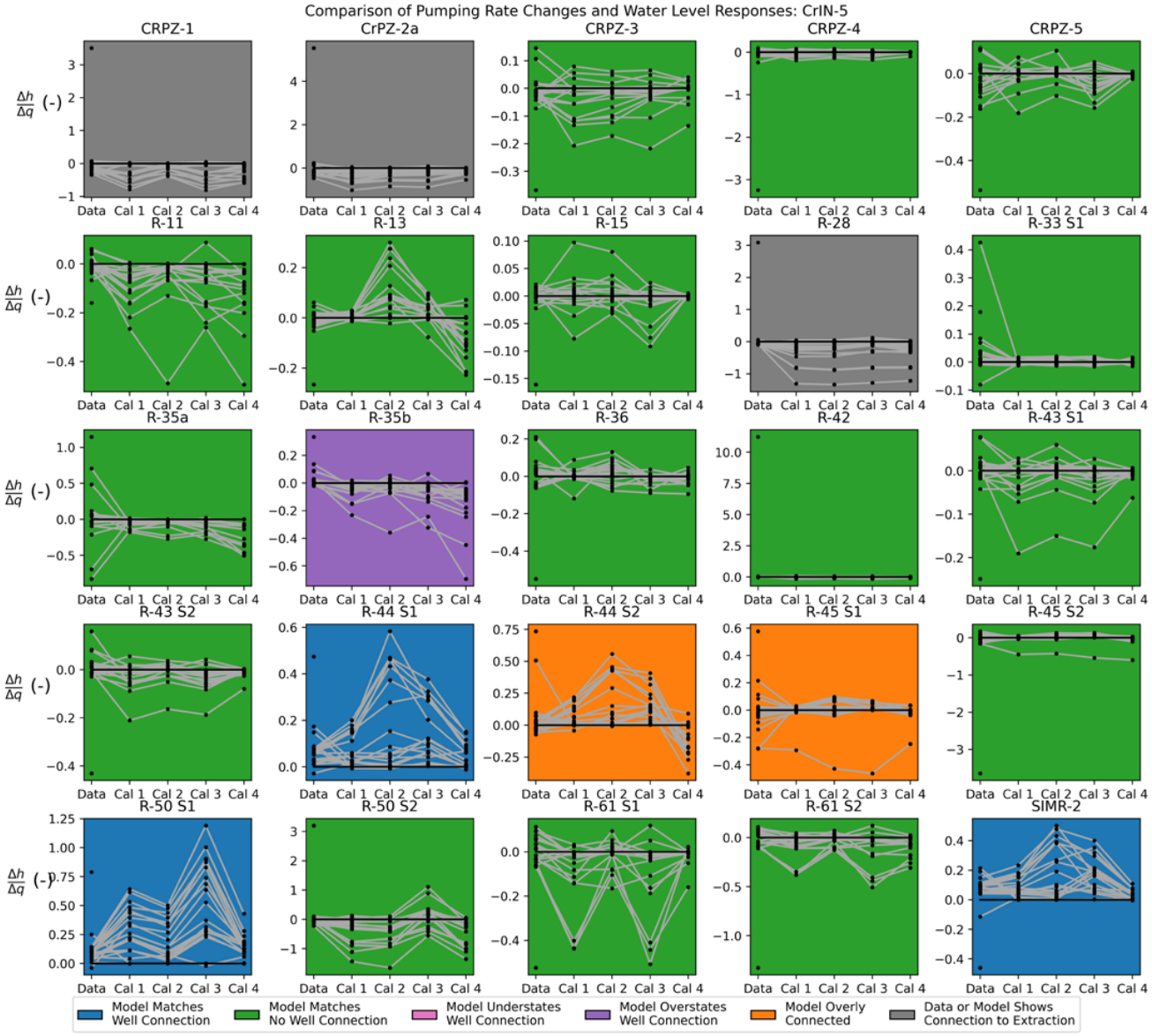

Figure A12. Ratio of the standardized water level response (Δh) to the standardized pumping rate change (Δq) for CrIN-5 well pairs. Faint grey lines connect the same event to allow comparison of each individual event among the data and each of calibrations 1 through 4. The $\Delta h/\Delta q$ ratio is unitless because Δq and Δh are normalized by their respective standard deviations.

Appendix B. Analysis of Residuals

Model diagnostic plots are used to assess whether the assumption of normal independent, identically distributed (iid) errors with constant variance is adequate: histograms of residuals are presented in "d" Figures A13–A18 and residuals against fitted values are plotted in "e" Figures A19–A24. For some pairs, a pattern in the residual plots indicates that the assumptions behind the statistical models are not met. In some cases, the variance in the residuals changes with the fitted value. This is to be expected because the data are a conglomeration of field measurements and results from several models, not random samples. The effect of the assumption violation is not immediately clear, though it is expected to lead to slightly larger *p*-values, which would have the consequence of identifying fewer significant connections in the data-only model. In the combined model that includes data and model results, inflated *p*-values would lead to more inferred matches.

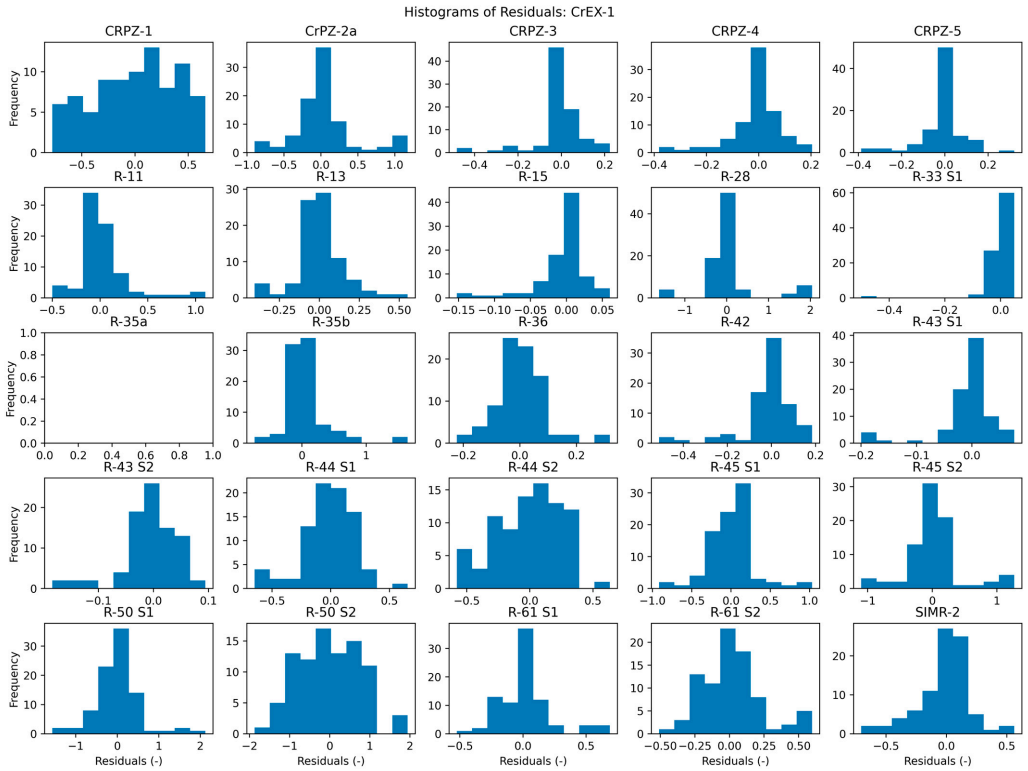

Figure A13. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrEX-1 well pairs. R-35a is excluded from the CrEX-1 analysis due to a low sample size (<15 events). Residuals are unitless because the predicted values are water levels normalized by standard deviation.

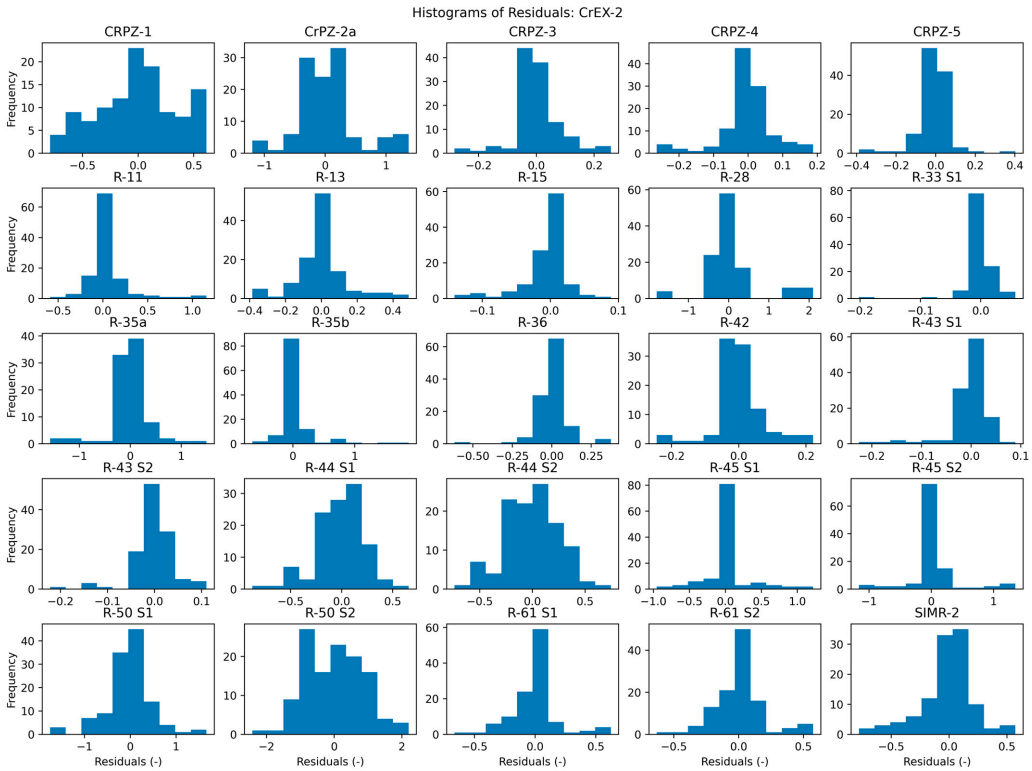

Figure A14. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrEX-2 well pairs. Residuals are unitless because the predicted values are water levels normalized by standard deviation.

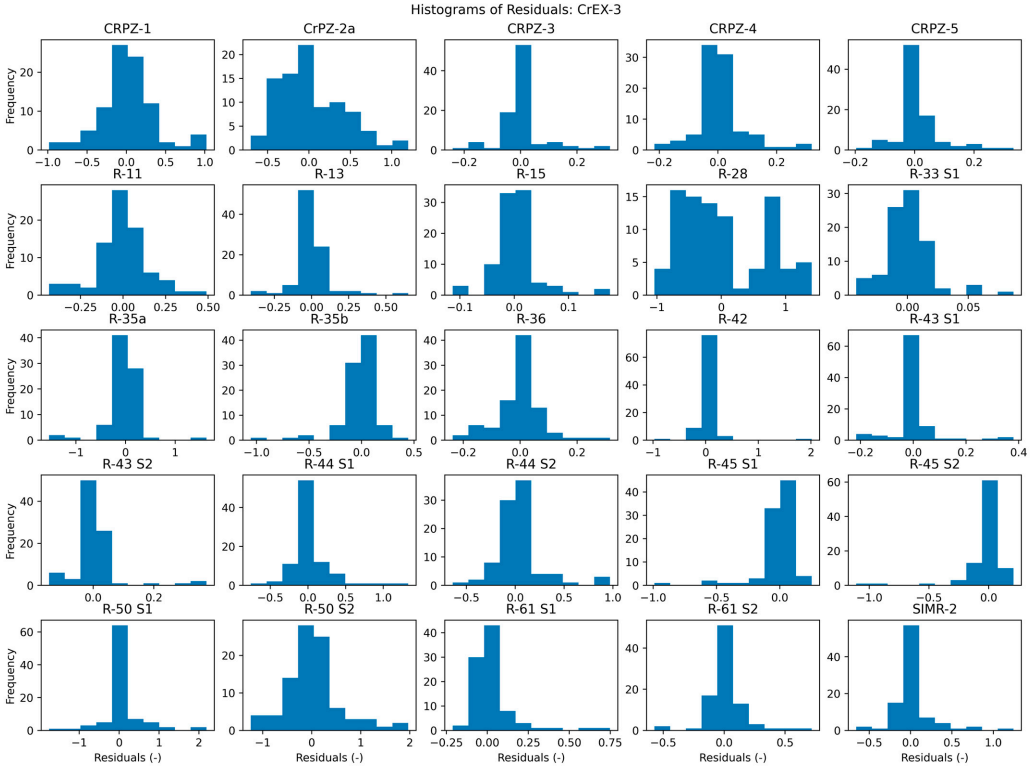

Figure A15. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrEX-3 well pairs. Residuals are unitless because the predicted values are water levels normalized by standard deviation.

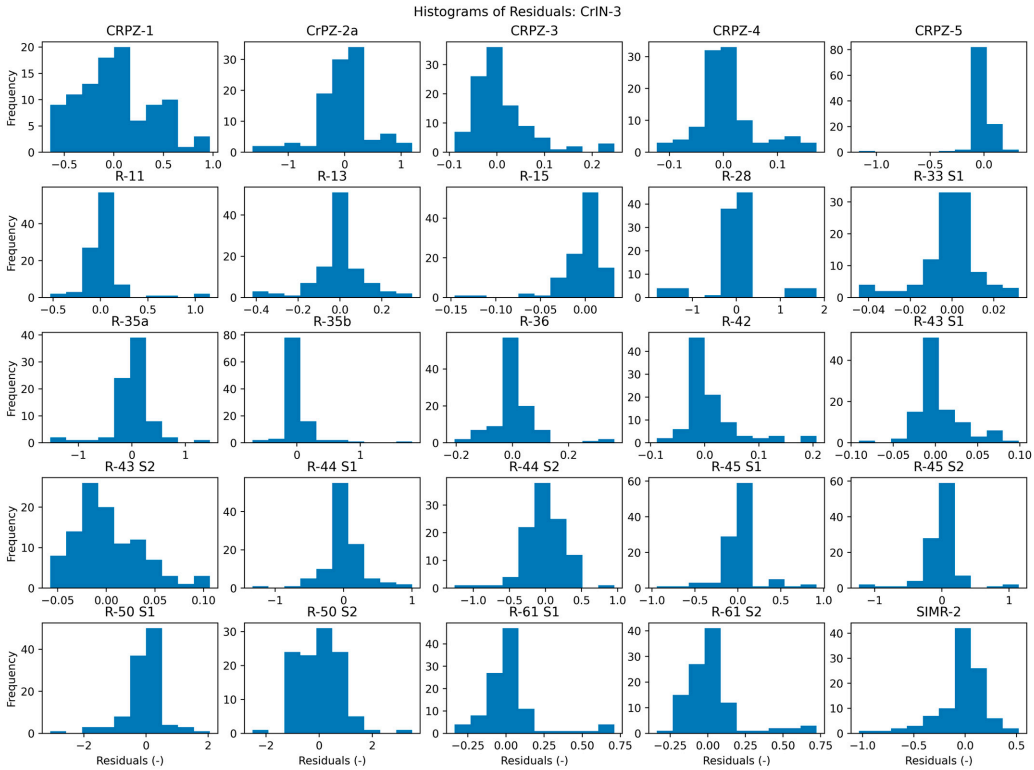

Figure A16. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrIN-3 well pairs. Residuals are unitless because the predicted values are water levels normalized by standard deviation.

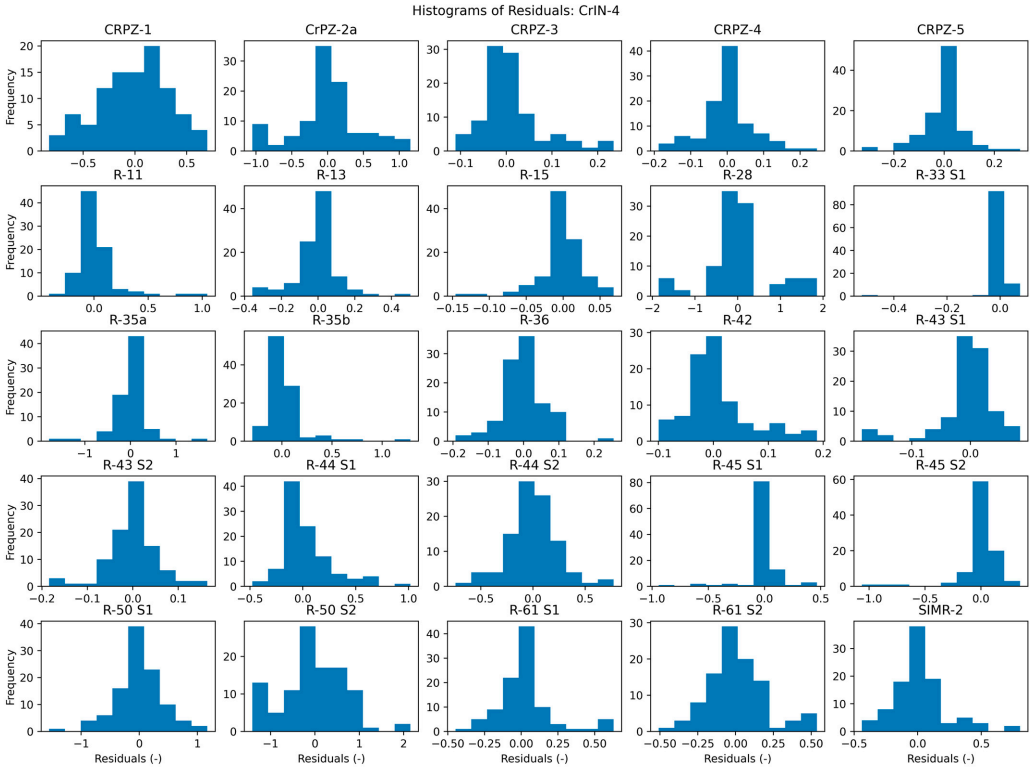

Figure A17. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrIN-4 well pairs. Residuals are unitless because the predicted values are water levels normalized by standard deviation.

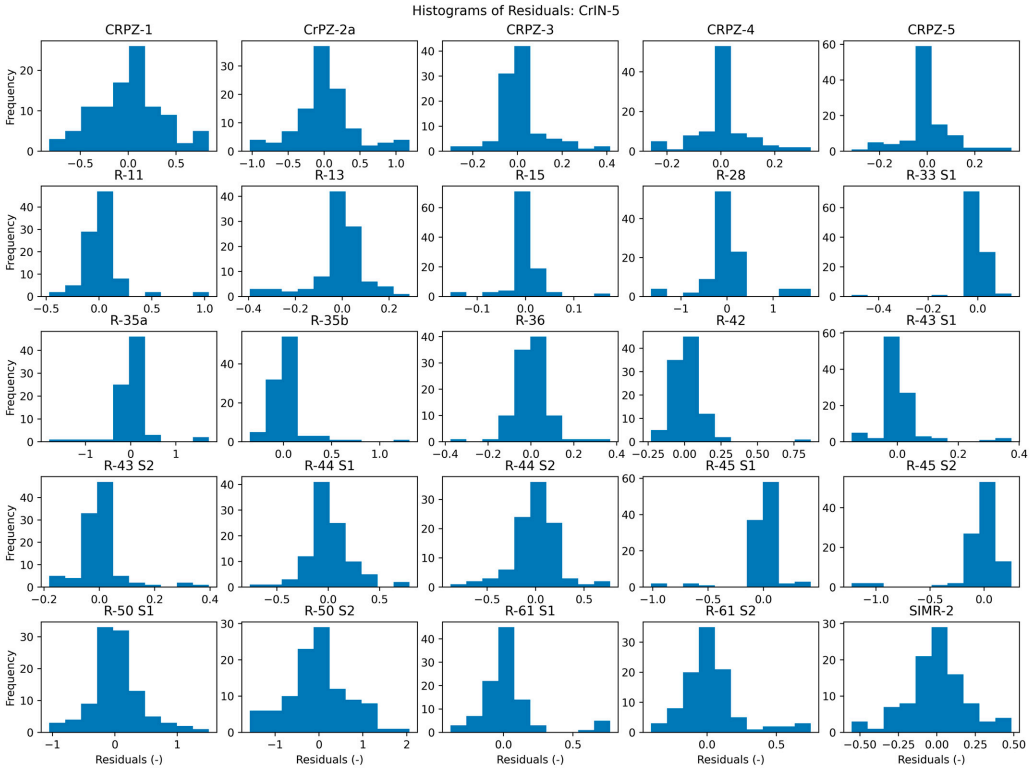

Figure A18. Histograms of residuals for Huber regressions of pumping rate changes and monitoring well responses for CrIN-5 well pairs. Residuals are unitless because the predicted values are water levels normalized by standard deviation.

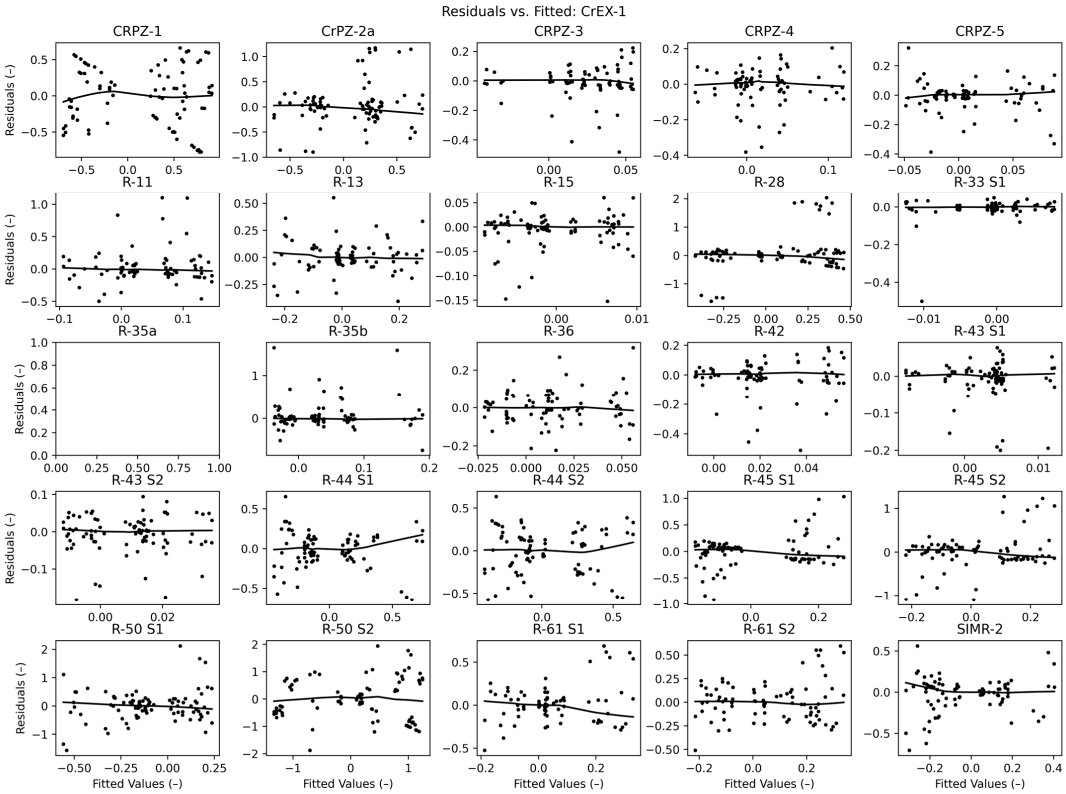

Figure A19. Residuals against fitted values for the CrEX-1 well pairs. R-35a is excluded from the CrEX-1 analysis due to a low sample size (<15 events). Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

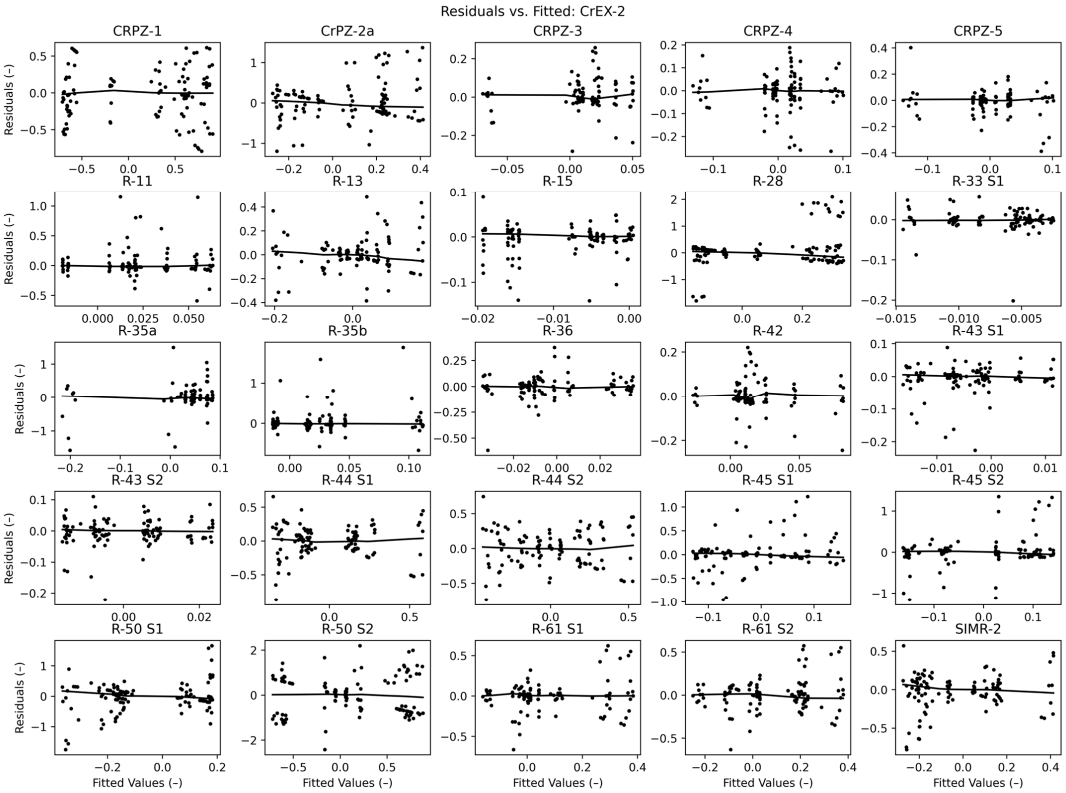

Figure A20. Residuals against fitted values for the CrEX-2 well pairs. Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

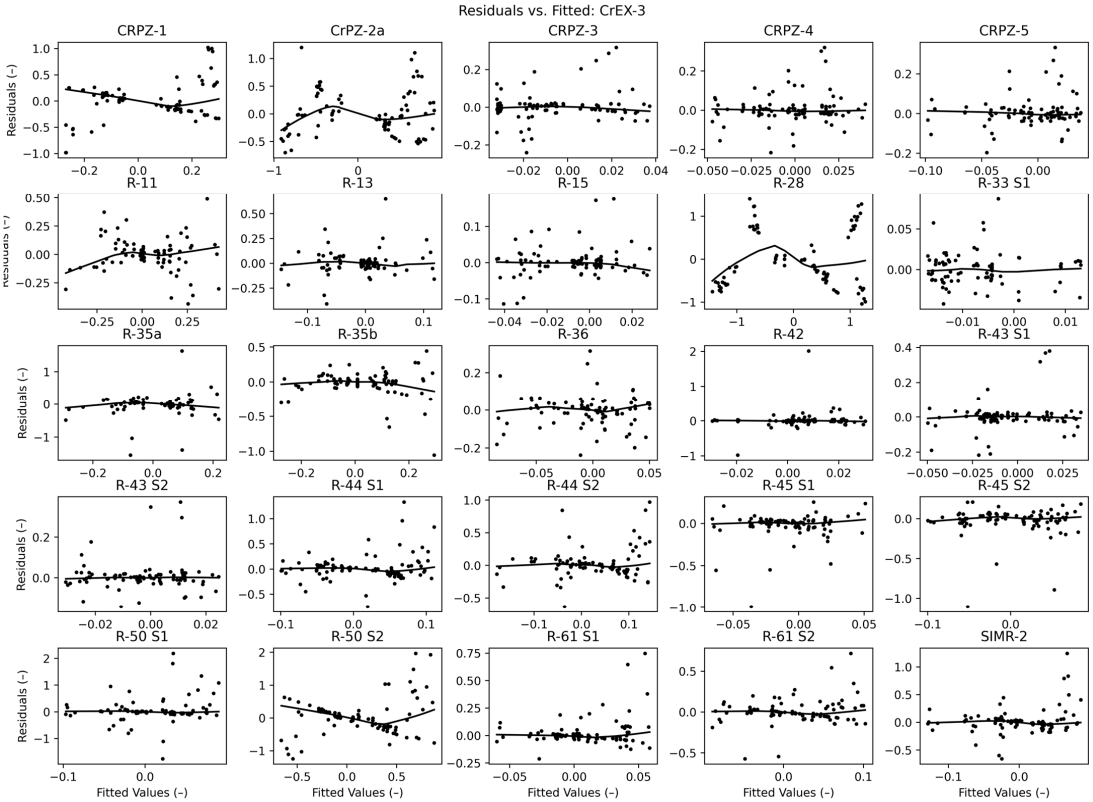

Figure A21. Residuals against fitted values for the CrEX-3 well pairs. Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

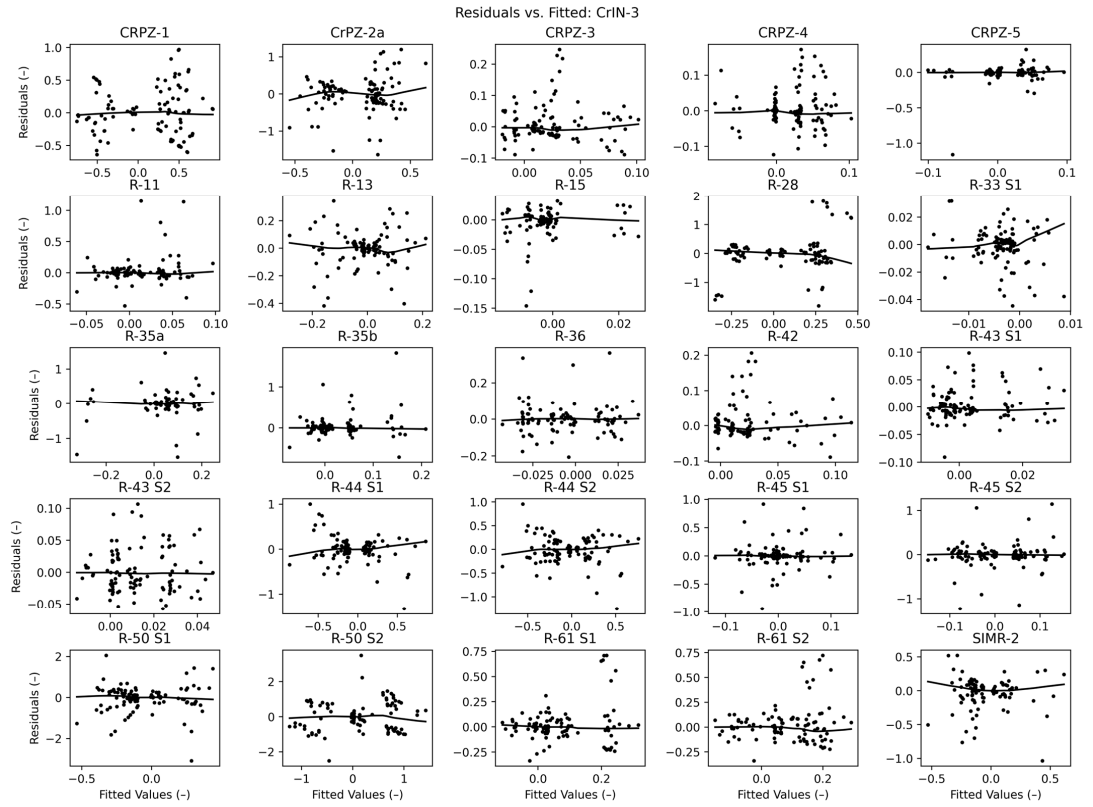

Figure A22. Residuals against fitted values for the CrIN-3 well pairs. Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

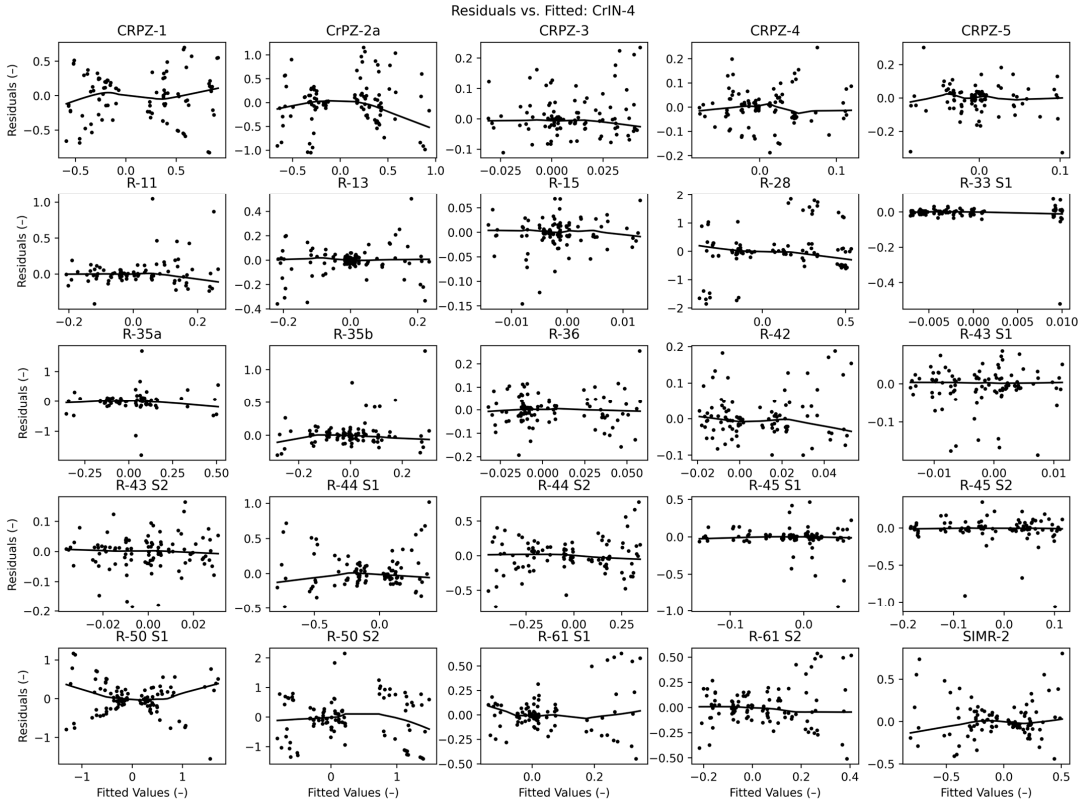

Figure A23. Residuals against fitted values for the CrIN-4 well pairs. Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

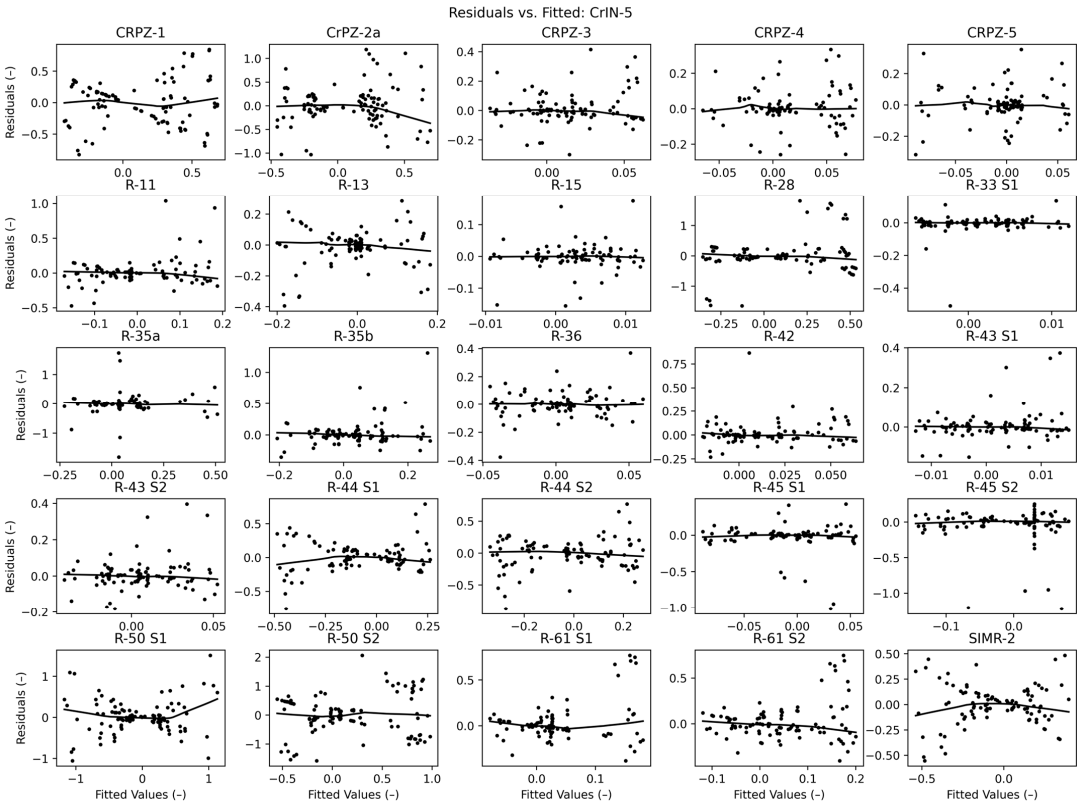

Figure A24. Residuals against fitted values for the CrIN-5 well pairs. Fitted values and residuals are unitless because the predicted values are water levels normalized by standard deviation.

Appendix C. Complete Set of Well Connection Maps

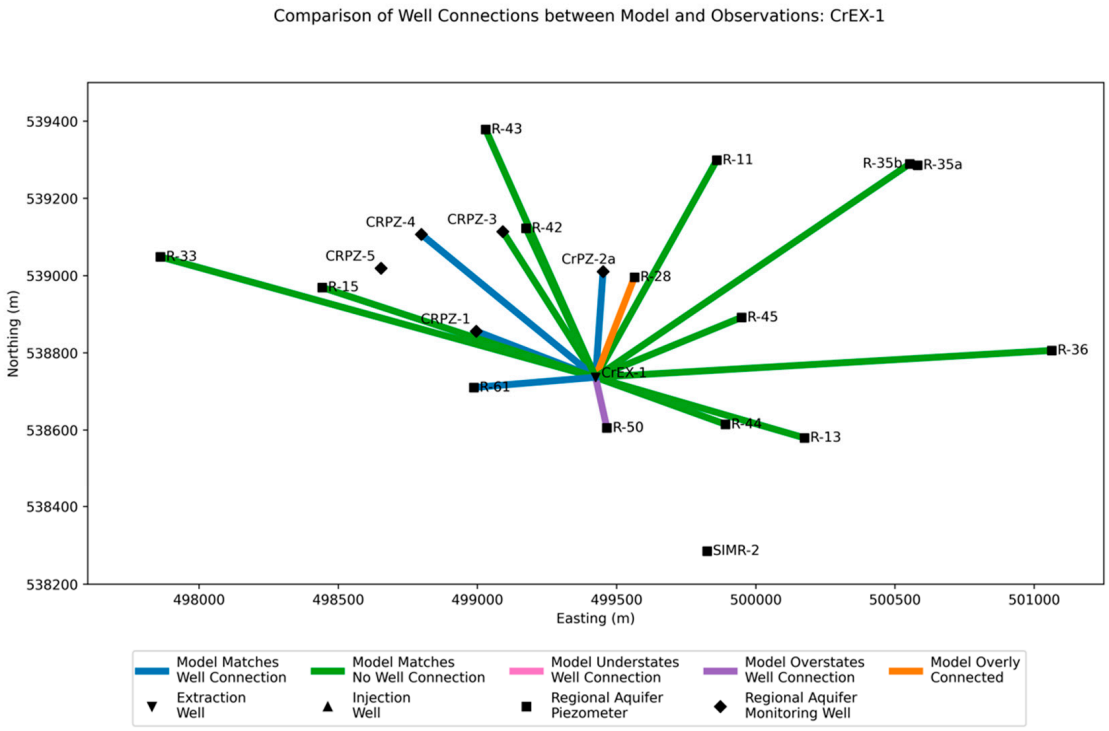

Figure A25. Map-based representation of connection classifications for CrEX-1 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

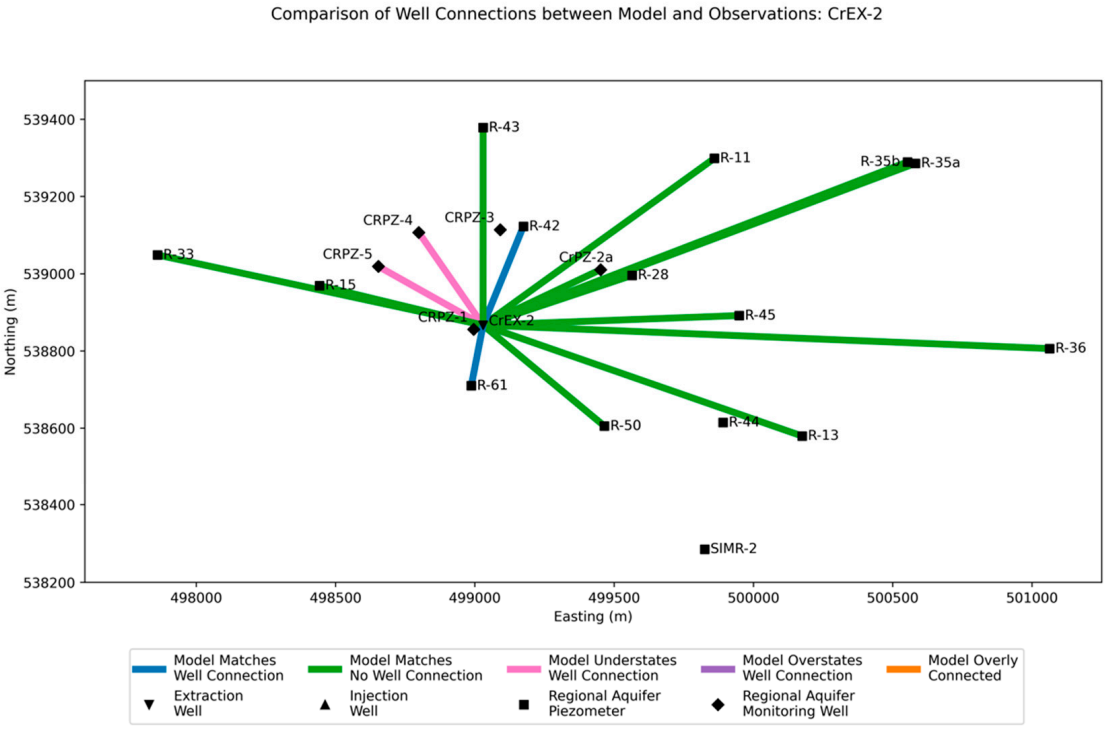

Figure A26. Map-based representation of connection classifications for CrEX-2 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

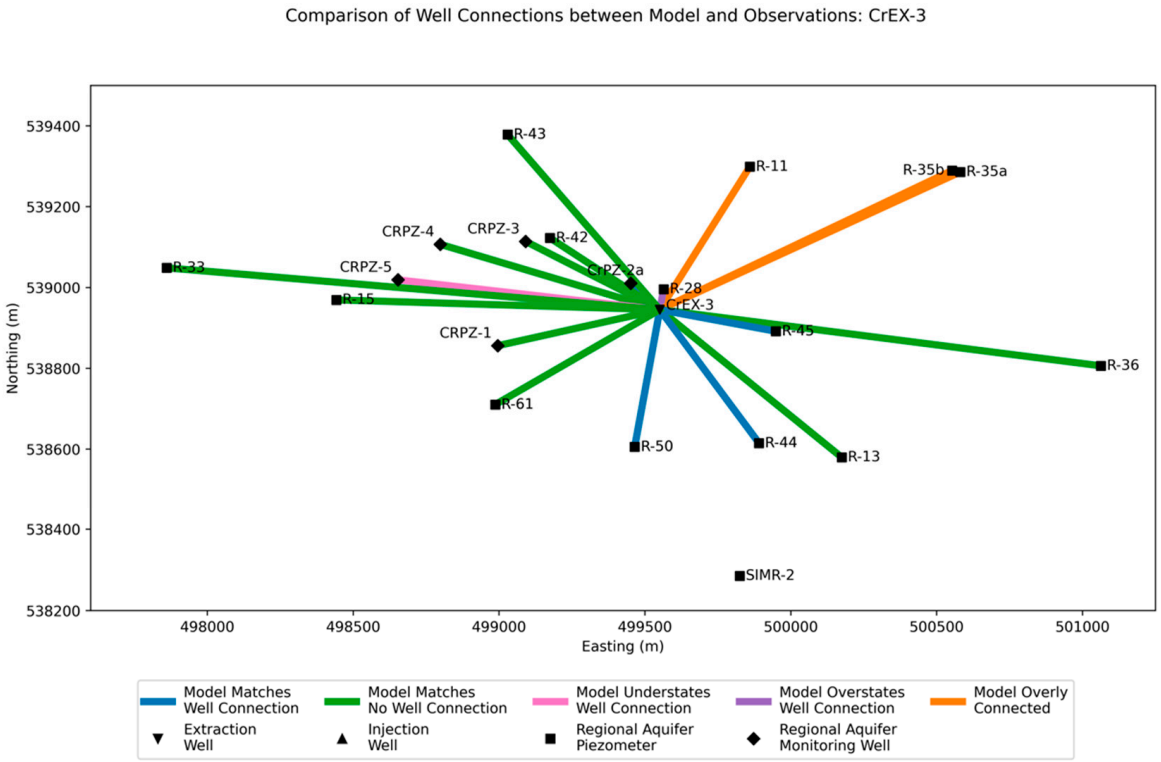

Figure A27. Map-based representation of connection classifications for CrEX-3 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

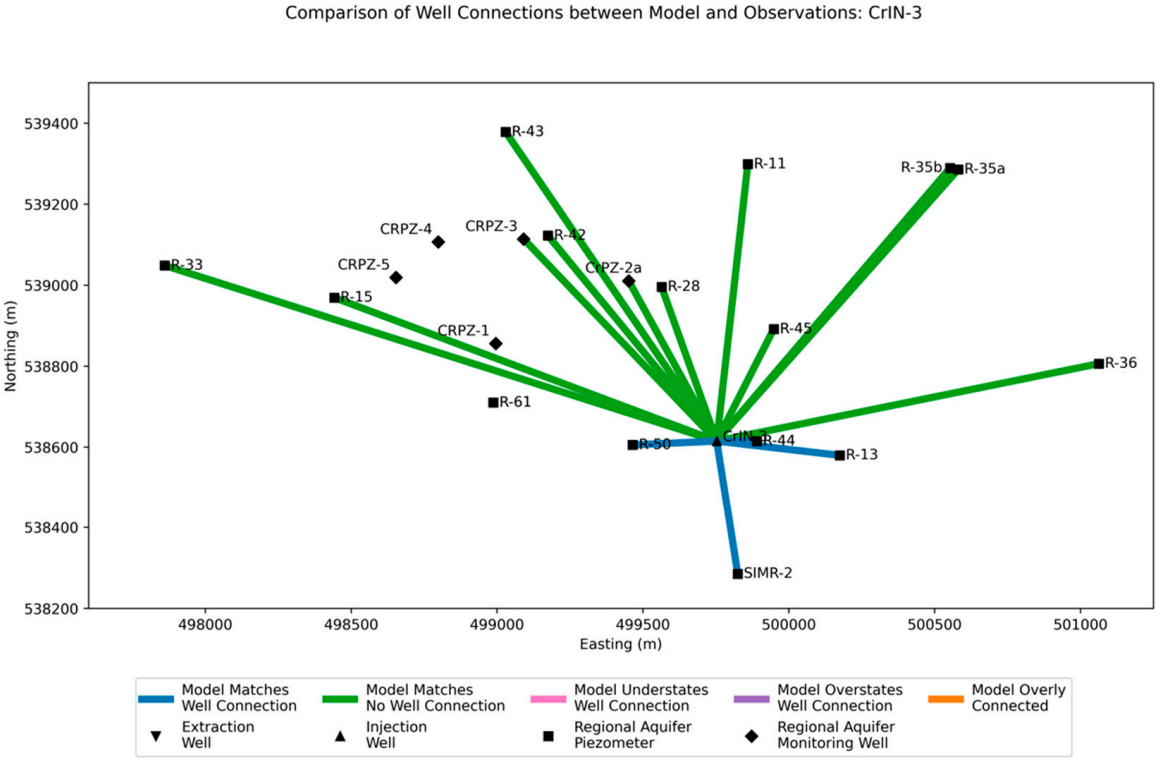

Figure A28. Map-based representation of connection classifications for CrIN-3 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

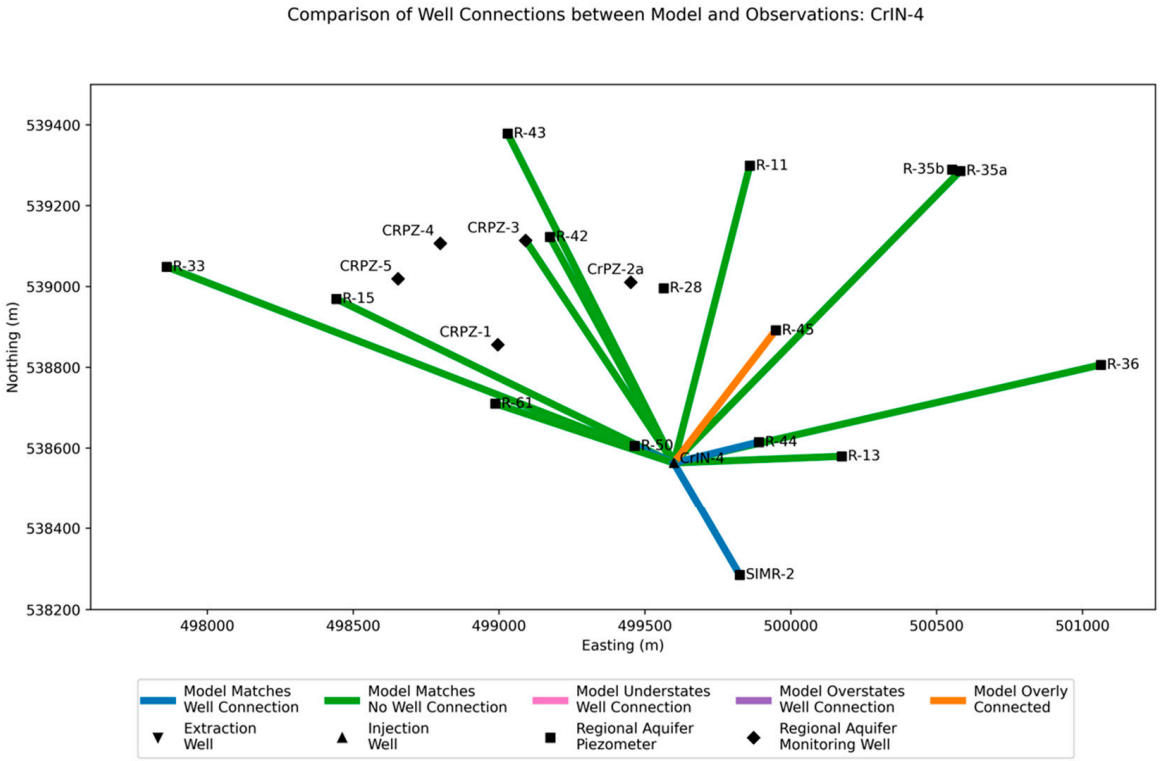

Figure A29. Map-based representation of connection classifications for CrIN-4 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

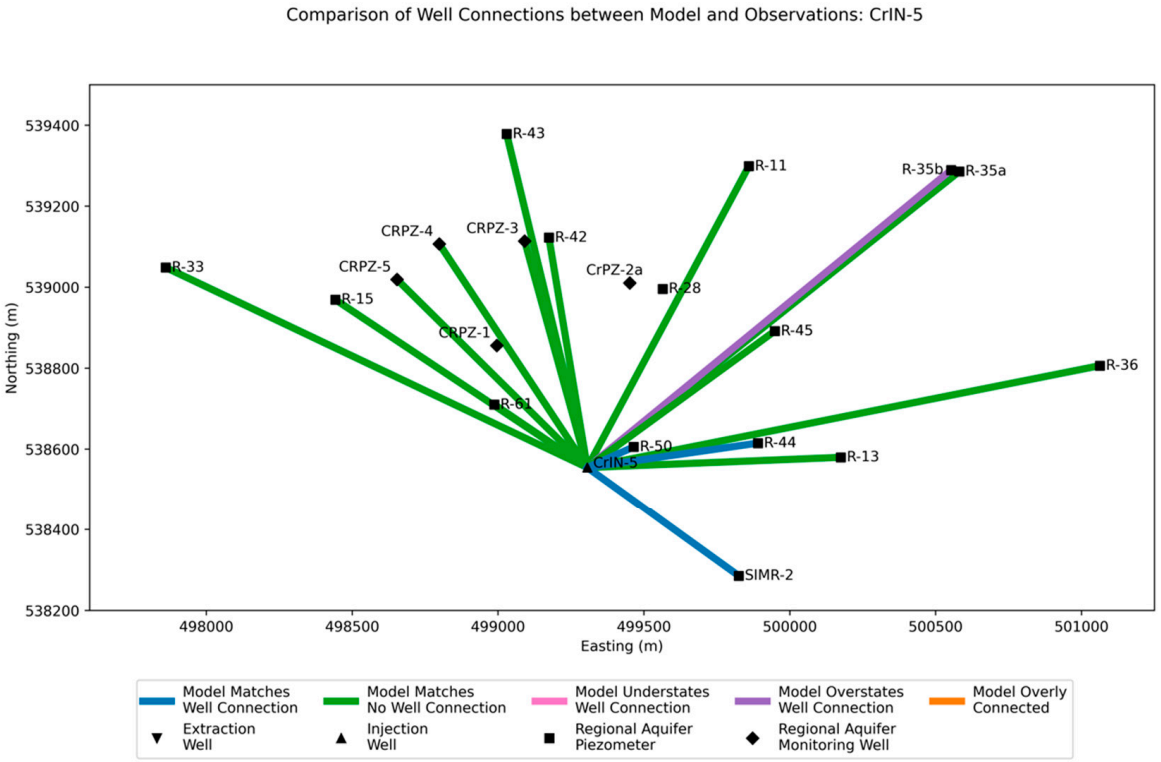

Figure A30. Map-based representation of connection classifications for CrIN-5 well pairs. Line colors summarize results across calibrations based on the most common outcome for the monitoring well to clarify the broad behavior of the model with respect to the data.

References

1. Finsterle, S.; Lanyon, B. Pragmatic Validation of Numerical Models Used for the Assessment of Radioactive Waste Repositories: A Perspective. *Energies* **2022**, *15*, 3585. [CrossRef]
2. Hassan, A.E. Groundwater Model Validation. Desert Research Institute, Nevada System of Higher Education, DRI/Bk/Chpt 2006. Available online: https://www.osti.gov/biblio/876739 (accessed on 24 November 2023).
3. Hassan, A. *A Validation Process for the Groundwater Flow and Transport Model of the Faultless Nuclear Test at Central Nevada Test Area*; DOE/NV/13609-27; Desert Research Institute, University and Community College System of Nevada (US): Reno, NV, USA, 2003; p. 45197. [CrossRef]
4. Hassan, A.E. Validation of Numerical Ground Water Models Used to Guide Decision Making. *Groundwater* **2004**, *42*, 277–290. [CrossRef] [PubMed]
5. Eisenberg, N.; Federline, M.; Sagar, B.; Wittmeyer, G. Model Validation From A Regulatory Perspective: A Summary. In *Validation Through Model Testing*; U.S. Nuclear Regulatory Commission: Washington, DC, USA, 1995.
6. Oreskes, N.; Shrader-Frechette, K.; Belitz, K. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* **1994**, *263*, 641–646. [CrossRef] [PubMed]
7. Anderson, M.P.; Woessner, W.W. The role of the postaudit in model validation. *Adv. Water Resour.* **1992**, *15*, 167–173. [CrossRef]
8. Flavelle, P. A quantitative measure of model validation and its potential use for regulatory purposes. *Adv. Water Resour.* **1992**, *15*, 5–13. [CrossRef]
9. Davis, P.A.; Olague, N.E.; Goodrich, M.T. *Approaches for the Validation of Models Used for Performance Assessment of High-Level Nuclear Waste Repositories*; NUREG/CR-5537; SAND-90-0575; Nuclear Regulatory Commission: Washington, DC, USA; Division of High-Level Waste Management, Sandia National Labs.: Albuquerque, NM, USA, 1991. Available online: https://www.osti.gov/biblio/5914752 (accessed on 24 November 2023).
10. Voss, C.F. *A Proposed Methodology for Validating Performance Assessment Models for the DOE Office of Civilian Radioactive Waste Management Program*; American Nuclear Society: Westmont, IL, USA, 1990.
11. Doherty, J. *Calibration and Uncertainty Analysis for Complex Environmental Models*; Watermark Numerical Computing: Brisbane, Australia, 2015; ISBN 978-0-9943786-0-6.
12. Illman, W.A.; Zhu, J.; Craig, A.J.; Yin, D. Comparison of aquifer characterization approaches through steady state groundwater model validation: A controlled laboratory sandbox study. *Water Resour. Res.* **2010**, *46*. [CrossRef]
13. Wikle, C.K.; Zammit-Mangion, A.; Cressie, N. *Spatio-Temporal Statistics with R*; Chapman and Hall/CRC: New York, NY, USA, 2019. [CrossRef]
14. Asadi, A.; Adhikari, K. Minimizing Errors in the Prediction of Water Levels Using Kriging Technique in Residuals of the Groundwater Model. *Water* **2022**, *14*, 426. [CrossRef]
15. Klaas, D.K.S.Y.; Imteaz, M.A.; Sudiayem, I.; Klaas, E.M.E.; Klaas, E.C.M. Assessing climate changes impacts on tropical karst catchment: Implications on groundwater resource sustainability and management strategies. *J. Hydrol.* **2020**, *582*, 124426. [CrossRef]
16. de Graaf, I.E.; van Beek, R.L.; Gleeson, T.; Moosdorf, N.; Schmitz, O.; Sutanudjaja, E.H.; Bierkens, M.F. A global-scale two-layer transient groundwater model: Development and application to groundwater depletion. *Adv. Water Resour.* **2017**, *102*, 53–67. [CrossRef]
17. Hong, N.; Hama, T.; Suenaga, Y.; Huang, X.; Ito, H.; Kawagoshi, Y. Simplified Lumped Groundwater Model to Simulate Nitrate Concentration Dynamics. *J. Hydrol. Eng.* **2017**, *22*. [CrossRef]
18. Xu, T.; Valocchi, A.J. Data-driven methods to improve baseflow prediction of a regional groundwater model. *Comput. Geosci.* **2015**, *85*, 124–136. [CrossRef]
19. Fienen, M.N.; Masterson, J.P.; Plant, N.G.; Gutierrez, B.T.; Thieler, E.R. Bridging groundwater models and decision support with a Bayesian network. *Water Resour. Res.* **2013**, *49*, 6459–6473. [CrossRef]
20. Marcot, B.G. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecol. Model.* **2012**, *230*, 50–62. [CrossRef]
21. White, J.T. A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environ. Model. Softw.* **2018**, *109*, 191–201. [CrossRef]
22. White, J.T.; Hunt, R.J.; Fienen, M.N.; Doherty, J.E. *Approaches to Highly Parameterized Inversion: PEST++ Version 5, a Software Suite for Parameter Estimation, Uncertainty Analysis, Management Optimization and Sensitivity Analysis*; Techniques and Methods 7-C26; U.S. Geological Survey: Reston, VA, USA, 2020. [CrossRef]
23. Doherty, J. Ground Water Model Calibration Using Pilot Points and Regularization. *Groundwater* **2003**, *41*, 170–177. [CrossRef] [PubMed]
24. Doherty, J.E.; Hunt, R.J. *Approaches to Highly Parameterized Inversion—A Guide to Using PEST for Groundwater-Model Calibration*; U.S. Geological Survey Scientific Investigations Report 2010–5169; U.S. Geological Survey: Reston, VA, USA, 2010. Available online: https://pubs.usgs.gov/publication/sir20105169 (accessed on 28 November 2023).
25. Luis, S.J.; McLaughlin, D. A stochastic approach to model validation. *Adv. Water Resour.* **1992**, *15*, 15–32. [CrossRef]
26. Harp, D.R.; Vesselinov, V.V. Identification of Pumping Influences in Long-Term Water Level Fluctuations. *Groundwater* **2011**, *49*, 403–414. [CrossRef] [PubMed]

27. The Relation Between the Lowering of the Piezometric Surface and the Rate and Duration of Discharge of a Well Using Ground-Water Storage. Eos, Transactions American Geophysical Union. 1935. Available online: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/TR016i002p00519 (accessed on 28 November 2023).
28. Los Alamos National Laboratory. *Fate and Transport Investigations Update for Chromium Contamination from Sandia Canyon*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2008.
29. Los Alamos National Laboratory. *Investigation Report for Sandia Canyon*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2009.
30. Los Alamos National Laboratory. *Phase II Investigation Report for Sandia Canyon*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2012.
31. Los Alamos National Laboratory. *Interim Measures Work Plan for Chromium Plume Control*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2015.
32. Los Alamos National Laboratory. *Chromium Plume Control Interim Measure Performance Monitoring Work Plan*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2018.
33. Newport News Nuclear BWXT-Los Alamos, LLC (N3B). *Assessment Report for the Evaluation of Conditions in the Regional Aquifer Around Well R-70*; Newport News Nuclear BWXT-Los Alamos, LLC (N3B): Los Alamos, NM, USA, 2021.
34. Newport News Nuclear BWXT-Los Alamos, LLC (N3B). *Annual Progress Report on Chromium Plume Control Interim Measure Performance, April 2022 Through March 2023*; Newport News Nuclear BWXT-Los Alamos, LLC (N3B): Los Alamos, NM, USA, 2023.
35. Neptune. *Chromium Interim Measure Capture Zone Analysis*; Neptune and Company Inc.: Los Alamos, NM, USA, 2023.
36. Neptune. *Chromium Model: Calibrated with Uncertainty Through 2022*; Neptune and Company Inc.: Los Alamos, NM, USA, 2023.
37. McLin, S.G. *Analyses of the PM-4 Aquifer Test Using Multiple Observation Wells*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2006.
38. Robinson, B.A.; Collins, K.; Simmons, A.M. *Los Alamos National Laboratory's Hydrogeologic Studies of the Pajarito Plateau: A Synthesis of Hydrogeologic Workplan Activities (1998–2004)*; Los Alamos National Laboratory: Los Alamos, NM, USA, 2005.
39. Neuman, S.P. Theory of flow in unconfined aquifers considering delayed response of the water table. *Water Resour. Res.* **1972**, *8*, 1031–1045. [CrossRef]
40. Hantush, M.S. Analysis of data from pumping tests in leaky aquifers. *Eos Trans. Am. Geophys. Union* **1956**, *37*, 702–714. [CrossRef]
41. Thiem, G. *Hydrologische Methoden*; J.M. Gebhardt: Leipzig, Germany, 1906.
42. Dupuit, J.É. *Études Théoriques et Pratiques Sur le Mouvement des Eaux Dans Les Canaux Découverts et à Travers Les Terrains Perméabls: Avec des Considérations Relatives au Régime des Grandes Eaux, au Débouché à Leur Donner, et à la Marche des Alluvions Dans les Rivières à Fond Mobile*; Dunod: Paris, France, 1863.
43. Huber, P.J. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Stat.* **1973**, *1*, 799–821. [CrossRef]
44. Huber, P.J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.