submitted to *Geophys. J. Int.*

A semblance measure for model comparison

Michael Commer

SUMMARY

Algorithmic and computational advances have made it possible that geophysical survey and earth model design can be aided by many systematic trial inverse-modeling runs with synthetic data. Such may for example come up in machine-learning approaches. Automated image appraisal pertaining to such applications will involve common statistical tests for goodness of data fit as a primary evaluation method. However, solution non-uniqueness may render multiple images equivalent in terms of their data fit, requiring secondary categorizers. A logical choice for classifying synthetic-imaging results quantifies the goodness of model fit where a known reference model replaces the observational input. The task of model inter-comparison in terms of measuring the resemblance to the reference model poses challenges to common distance-based metrics like root mean square error and mean absolute error. First, distance-based metrics can introduce spurious contributions when smooth models with fuzzy target contours are to be compared against a sharp reference. Second, large differences due to parameter-estimation overshoots can dominate distance metrics. The remedy proposed here is referred to as semblance and is based on the idea of logistic functions, where a binary dependent variable adds non-zero or zero accumulation terms for the, respectively, passing or failing of preset target thresholds. This classifying approach is amenable to an objective where model feature recognition is primary. Numerical comparisons to distance-based metrics provide evidence for the advantages of the semblance in view of this objective. Geophysical imaging in conjunction with machine-learning is seen as a benefitting upcoming application area.

Keywords: Electrical resistivity tomography (ERT), Hydrogeophysics, Image processing, Numerical modelling.

1 INTRODUCTION

Algorithmic and computational improvements to geophysical modeling capabilities have rendered geophysical forward- and inverse-modeling feasible within reasonable timeframes. These developments are likely to benefit applications that analyze multiple modeling outputs, which may come up in machine-learning contexts (Karpatne *et al.* 2019). Examples are multi-task learning for model optimization (Caruana 1993) or learning algorithms that contain iterative inverse problems (Kellman *et al.* 2020), that is, perform inversions in order to learn to invert (Putzky & Welling 2019).

Improved simulation capabilities have also been regarded as an instigator to more sophisticated statistical experimental geophysical survey design (Maurer *et al.* 2010). Survey design estimates best field configuration parameters that maximize the subsurface resolution ability of the resulting data set. If cast into an iterative inverse problem, one would repeat an underlying synthetic-data inversion procedure until an economically feasible survey configuration reproduces a subsurface target of interest. Hence, this has also been referred to as a macro-optimization problem (Maurer *et al.* 2010).

With these considerations in mind, it is the author's belief that the advent of deep-learning methods applied to survey design (e.g., Nakayama et al. 2019), in combination with capable simulators, will make macro-optimization problems with many synthetic images to be analyzed an increasingly common application area. Note that in the following, I will refer to image as the instance of a subsurface model, quantifying a discrete set of property parameters and abbreviated by the model vector $\mathbf{m}$ of size M. Within a macro-optimization framework, an image $\mathbf{m}$ can be the product of some model-generating procedure like inversion, where the term imaging shall be used synonymously to inversion.

The evaluation of a possibly large number of images leads to the question of appropriate goodness-of-fit measures (Huber-Carol *et al.* 2002). Goodness of fit quantifies accuracy through the departure of a model's prediction from some given reference, the latter typically given by a set of observations that are deemed reliable (e.g., Willmott *et al.* 2005). In the case of synthetic-data inversions, the reference data is produced by a reference model. The latter will also be called true model in the following and abbreviated by $\tilde{\mathbf{m}}$. Within the aforementioned contexts where the numerical rating of multiple trial images steers some optimization process, the decision-making ability of goodness-of-fit measures becomes paramount.

Goodness of fit as the core for scalar-based modeling performance metrics is usually linked to the concept of distance (Willmott *et al.* 1985; Deza & Deza 2009). Distance metrics quantify data goodness of fit through calculating the discrepancy between a reference (true, or observed) set $\mathbf{d} = (\tilde{d}_1, \ldots, \tilde{d}_N)$ of N data points and a predicted (or estimated) set $F(\mathbf{m}) = (d_1, \ldots, d_N)$, where F represents a forward-modeling operator. The discrepancy is the error vector $\mathbf{e} = \mathbf{d} - \tilde{\mathbf{d}}$ with the individual components $e_i = d_i - \tilde{d}_i$ $(i = 1, \ldots, N)$. In a review article, Botchkarev (2019) categorized

performance metrics used in machine learning, regression, and forecasting, which involved the analysis of approximately 500 journal articles and contained a list of over 50 common metric formulas. It is noteworthy that every reported metric formula is based on the error term $e_i = d_i - \tilde{d}_i$. Also, the two most prominent errors are the mean square error (RMSE) and the mean absolute error (MAE) (Willmott *et al.* 1985; Willmott & Matsuura 2005; Chai & Draxler 2014), also referred to as L2 and L1 norm. This contribution is motivated by my perception that upcoming application areas that require the evaluation of many images merit revisiting the adequacy of such traditional fit-measuring metrics. The reason stems from two common issues hampering geophysical imaging, solution non-uniquenss and model smoothness.

Distance-based model evaluation criteria that quantify data goodness of fit involve the calculation of $\mathbf{e} = F(\mathbf{m}) - F(\tilde{\mathbf{m}})$ in synthetic-data imaging. However, the problem of non-uniqueness (e.g., Jackson 1979), a commonality in geophysical imaging with overparameterized models, can render two different images $\mathbf{m}_1$ and $\mathbf{m}_2$ equivalent in terms of their data fit. Hence, the data predictions $F(\mathbf{m}_1) = F(\mathbf{m}_2)$, if used solely, would be an indifferent criteria for rating $\mathbf{m}_1$ against $\mathbf{m}_2$. Given that the reference $\tilde{\mathbf{m}}$ is known, the logical choice for an alternative or additional criteria for image intercomparison quantifies model goodness of fit. Then, distance metrics will contain the core $e_i = m_i - \tilde{m}_i$ ($i = 1, \ldots, M$) for quantifying the discrepancy between a model prediction $\mathbf{m}$ and its true state $\tilde{\mathbf{m}}$, which leads to the second issue, model smoothness.

Subsurface model predictions produced by geophysical imaging are typically smooth, either due to solution-stabilizing functionals (e.g. Portniaguine & Zhdanov 1999; Menke 2018), or simply due to the instrumentally limited resolution capability of potential-field and low-frequency (electromagnetic or seismic) methods. With limited image resolution, model evaluation usually reverts to focusing on local regions of anomalous trends, expressed for example as seismic low-velocity zones, electrically conductive faults, or hydrological (high-permeability) preferential fluid paths. The main motivation for this work is that in view of more fuzzy images and a trend-focused evaluation objective, L2 (and L1) types of metrics underperform; in other words, they lack the capacity of ”zooming” in on model features of interest.

To the best of my knowledge, as also indicated by the in-depth review of Botchkarev (2019), the overwhelming number of model evaluations employ RMSE, MAE, or variants thereof. Geostatistical inversion in hydrogeological studies (e.g., Kitanidis 1997) appears to be an application area where RMSE and MAE are often employed concurrently to quantify model goodness of fit (e.g., Hughson & Yeh 2000; Yeh & Liu 2000; Kowalsky *et al.* 2012); perhaps owing to the fact that L2 and L1 summarize variance and bias, respectively, of true-versus-estimated scatter plots, the latter a practical way of visualizing model fits.

Smooth images from geostatistical methods can exhibit a striking discordance between a visually perceived fit of local anomalies and the actual global L2 (or L1) norm. For example, Hughson & Yeh (2000) estimated a 3D volume of hydraulic conductivity from pressure data. Minimal L1 and L2 norms were reported for an inversion realization that employed the most input data (figure 2d). However, the corresponding 3D section (plate 1e) appears to underrepresent local field anomalies when comparing to the image with maximal norm and wider scatter plot (figure 2e and plate 1f). These observations support my claim that a feature- or trend-oriented model evaluation might produce a different ranking than global distance-based norms.

For the purpose of image inter-comparison that is driven by model-feature identification, I thus propose an alternative metric based on classifying functions, the latter enabled by step or sigmoidal function types. Used for classification problems in machine-learning, sigmoidal (or S) functions (Russell 2013) in principal facilitate replacing the accumulation of differences by delta-like function contributions. A classifying approach to model resemblance is more amenable to a softened objective that seeks to identify local model trends, instead of minimizing a global distance-based error norm. This is due to the overshoot-forgiving nature of the S function. Hence, with the evaluation objective shifting from global norm minimization to resemblance of features, the metric will be referred to as semblance and abbreviated by S.

Before introducing the semblance formulation in the next section, I want to emphasize that no actual inverse modeling is carried out in this work. The main focus, to be detailed in subsequent sections, is to highlight where the semblance metric is potentially superior over prominent distance-based candidates, RMSE and MAE. Given the breadth and complexity of learning types of algorithms, any exemplifying presentation within such a framework would go too much beyond this intended scope. Lastly, while potentially useful for a more targeted model analysis, the S metric should only be regarded as an additional flexible tool to be available among the many existing misfit measures.

2 METHODOLOGY

I refer to image as some representation of a known earth model. Being the product of geophysical imaging (synonymous to inverse modeling or inversion in this context) in the above mentioned application types, images usually exhibit some shortcomings, like loss of resolution or artifacts. This section is dedicated to the quantification of these shortcomings by means of a likeness-metric between image and the actual model.

The earth models considered here are represented as the vector $\mathbf{m}$ of size M specifying a discrete set of subsurface properties of interest. The vector $\mathbf{m}$ is defined over a gridded volume representing a modeling domain Ω, where the grid may be of type finite differences, finite elements, or other.

Demonstration examples below will involve images of electrical conductivity, σ, as the property type of interest. Note that electrical conductivity, measured in Sm^{-1}, will be used interchangeably with its inverse, electrical resistivity $\varrho = \frac{1}{\sigma}$, measured in $\Omega\mathrm{m}$.

It shall be reiterated that this section derives a semblance metric to be used solely for post-inversion image inter-comparison. While some of the presented model comparisons are framed within hypothetical inversion contexts, no actual inverse modeling is carried out. Some example model comparisons analyzed below originate from imaging results derived in earlier publications.

2.1 Common distance-based metrics

As a starting point, let us first recall two of the most common distance-based measures, because they will provide benchmarks for the presented metric comparisons. The first is the root mean square error (RMSE) (e.g., Willmott & Matsuura 2005) and is written here for the purpose of quantifying the difference between the prediction (image) $\mathbf{m}$ of a reference model $\tilde{\mathbf{m}}$,

$$\varepsilon_{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(m_i - \tilde{m}_i)^2}. \tag{1}$$

Any weighting of the individual difference terms is omitted, although such weighting could be useful if the certainty about the reference $\tilde{\mathbf{m}}$ varies spatially. The comparative image analysis below will employ a second metric, the mean absolute error (MAE),

$$\varepsilon_{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|m_i - \tilde{m}_i\right|. \tag{2}$$

RMSE and MAE are also referred to as absolute error norm L1 and mean-square error norm L2, respectively (e.g., Yeh & Liu 2000). There exists some controversy about RMSE versus MAE (e.g., Willmott & Matsuura 2005; Chai & Draxler 2014), as also reviewed by Botchkarev (2019). In short, RMSE can be more ambiguous than MAE as it does not describe average error alone and has other implications that can obscure a primarily averaging purpose. On the other hand, it is argued that RMSE is more in line with many statistical calculations.

Earth model properties like electrical resistivity typically exhibit a large range, often over several decades. Therefore, it is common to treat such properties in logarithmic space, leading to the logarithmic counterparts of eqs (1) and (2),

$$\varepsilon^{log}_{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\log(m_i) - \log(\tilde{m}_i)\right)^2}, \tag{3}$$

$$\varepsilon^{log}_{MAE} = \frac{1}{M}\sum_{i=1}^{M}\left|\log(m_i) - \log(\tilde{m}_i)\right|. \tag{4}$$

Employing eqs (1)-(4) as model-constraining objective functions in inversion contexts would have differing effects on an imaging outcome; hence, L1 or L2 types of measures or more elaborate variants may be constructed for their specific and predictable effects on the minimization process (e.g., Farquharson 2008). Similarly, for post-inversion model inter-comparison, different norms translate to different lenses for measuring discrepancies in certain model features of interest. However, in contrast to the unknown inversion target, the standalone numerical comparison to a fully known model offers the advantage of a much higher adaptability of the metric to the target. The semblance metric to be derived in the following seeks to exploit this advantage through threshold parameters.

2.2 Thresholds for model features: the indifference of difference measures

Consider a simple model with two electrical resistivity parameters illustrated in Fig. 1. The true model describes a target $\tilde{\mathbf{m}} = (\tilde{m}_1, \tilde{m}_2) = (30, 30)$ Ωm, which is less resistive than an initial model guess $\mathbf{m}^0 = (m_1, m_2) = (100, 100)$ Ωm. Preceding hydrogeological studies may dictate that in order for the target to qualify as sufficiently anomalous, here conductive, an image is expected to yield $m_i \leq 30$ Ωm, marked as anomaly threshold in Figure 1. One image may yield $\mathbf{m}^1 = (50, 35)$, another one $\mathbf{m}^2 = (10, 25)$. Exemplified for eq. (1), the model guesses produce the same error $\varepsilon^1_{RMSE} = \varepsilon^2_{RMSE} \approx 15$, indicating a 79 % improvement with respect to the initial guess $\varepsilon^0_{RMSE} = 70$.

The example attempts to highlight some inadequacy of difference measures when the objective is to discern the character of an anomaly, here the low-resistivity zone. Squared or absolute differences do not discriminate between positive and negative error contributions. Therefore, RMSE or MAE types of errors remain indifferent to the fact that both parameters of Image $\mathbf{m}^1$ fail the preset criteria $m_i \leq 30$ Ωm for a conductive anomaly while Image $\mathbf{m}^2$ fully passes. Note that a simple difference, i.e. one that retains the sign of the terms $(m_i - \tilde{m}_i)$, may also be inaccurate owing to potential cancellation of individual error contributions.

Motivated by this indifference of distance-based metrics, the subsequent section derives an alternative metric, to be called semblance and abbreviated by S. The metric basically counts those model parameters that meet the anomaly criteria $m_i \leq 30$, which makes it able to discern the improvement of Image $\mathbf{m}^2$ (all parameters qualify, thus $S^2 = 100$ %) over Image $\mathbf{m}^1$ (none qualifies, thus $S^1 = 0$ %). In principal, the accumulation of distance-based error contributions is replaced by an accumulation of non-zero function contributions. For the example of Fig. 1, the latter become non-zero when model parameters meet the criteria $m_i \leq 30$ Ωm.

2.3 Parameterizing model semblance through lower and upper bounds

Distance-based metrics like eqs (1)-(4) measure how far apart two models are in model space. The example of Fig. 1 suggests to shift this paradigm to something describing the semblance of two models in terms of distinct features of interest. Semblance may be more amenable to imaging that attempts to identify anomalous regions, like zones of low/high velocity, resistivity, density, etc., that stand out from some known (or assumed known) geologic background.

While there exist numerous ways, this work considers two types of anomaly-defining criteria. The first is referred to as boundary condition and is parameterized through a function $S \sim h^{bnd} = h^{bnd}(a,b)$ with lower and upper boundaries a, b that define an anomalous property range $[a,b]$. In a discrete implementation, the semblance (S) distinguishes between target matches $h^{bnd} = 1$ and mismatches $h^{bnd} = 0$ for a given model element m_i,

$$h^{bnd}(m_i,a,b) = \begin{cases} 1 & \text{for} \quad m_i \in [a,b], \quad \text{where } [a,b] \Rightarrow x \in \Omega \in \mathbb{R} : a \leq x \leq b \\ 0 & \text{for} \quad m_i \notin [a,b]. \end{cases} \tag{5}$$

Evaluating the similarity between the true model $\tilde{\mathbf{m}}$ and one of the images in Fig. 1 in terms of their conductive nature suggests to set $[a,b] = [0,30]$ Ωm, yielding a maximal semblance for Image $\mathbf{m}^2$.

Fig. 2a exemplifies h^{bnd} for a different scenario $[a,b] = [0.1,1.1]$. Consider a fluid injection example where one attempts to image the electrical conductivity of a freshwater plume within a saline and thus conductive background, the latter characterized by $\mathbf{m}^0 = 3$ Sm^{-1}. Preliminary hydrological studies may dictate an upper bound of b =1.1 Sm^{-1} for labeling grid elements as sufficiently resistive. For now, consider only the step-function representation of h^{bnd} in Fig. 2a. Imposing the boundaries $[a,b] = [0.1,1.1]$ Sm^{-1} produces $h^{bnd}(m_i,0.1,1.1) = 1$ for all elements σ_i with $0.1 \leq \sigma_i \leq 1.1$ Sm^{-1}, deeming them as anomalous. The lower bound of $a = 0.1$ would have the effect of penalizing images with resistive overshoots, that is, $h^{bnd} = 0$ for all $\sigma_i < 0.1$. An alternative choice $a = 0$ would include all (resistive) cells $\sigma_i \leq 1.1$.

Given the fact that geophysical modeling often involves the evaluation of time-series of images, the second anomaly-defining criteria introduced next involves relative property changes.

2.4 Parameterizing model semblance through bounds for relative changes

Let t=0 mark the beginning of some time-lapse data acquisition, like a fluid injection experiment. Then, the time $t > 0$ (after injection begin) defines the survey time at which one wants to evaluate an image $\mathbf{m}^t$ of the actual subsurface state $\tilde{\mathbf{m}}^t$. The objective is to compare models in terms of their time-lapse change relative to an undisturbed background (here pre-injection) state $\mathbf{m}^{t=0} = \mathbf{m}^0$, facilitated

by the relative-difference measures $\tilde{\Delta}$ and Δ,

$$\tilde{\Delta}_i = \frac{\tilde{m}_i^t - m_i^0}{m_i^0}, \tag{6}$$

$$\Delta_i = \frac{m_i^t - m_i^0}{m_i^0}. \tag{7}$$

For a given model grid element i, $\tilde{\Delta}_i$ measures the actual relative property change between $\mathbf{m}^0$ and $\tilde{\mathbf{m}}^t$ that occurs over the time t. Similarly, Δ_i is the relative change for a corresponding image $\mathbf{m}^t$. Note that both relative differences are calculated with respect to the same background model $\mathbf{m}^0$, despite the fact that an imaging method producing $\mathbf{m}^t$ may involve a different background as starting model.

Parameterizing a semblance function $S \sim h^{\Delta}$ for labeling a grid element's temporal property evolution as anomalous, incorporates eqs (6) and (7) in two discrete ways,

$$h_+^{\Delta}(\Delta_i^t, \Delta^{lim}) = \begin{cases} 1 & \text{for} \quad \Delta_i^t \geq \Delta^{lim} \\ 0 & \text{for} \quad \Delta_i^t < \Delta^{lim} \end{cases} \tag{8}$$

$$h_-^{\Delta}(\Delta_i^t, \Delta^{lim}) = \begin{cases} 1 & \text{for} \quad \Delta_i^t \leq \Delta^{lim} \\ 0 & \text{for} \quad \Delta_i^t > \Delta^{lim} \end{cases} \tag{9}$$

Each of these two cases is referred to as relative condition, where eq. (8) describes a relative property increase over time, while eq. (9) describes a decrease. Fig. 2b exemplifies both cases by the step-like functions denoted as h_+^{Δ} and h_-^{Δ}. In this example, the limit is set to $\Delta^{lim} = -0.15$. The aforementioned fluid injection case with a growing resistive plume as model feature of interest would employ the negative relative condition h_-^{Δ}, eq. (9). Hence, enforcing $\Delta_i^t \leq \Delta^{lim} = -0.15$ would label each grid cell σ_i^t with a time-lapse conductivity decrease of 15 % or more as anomalous.

While relative changes are described here in a temporal manner, the relative condition can of course also involve spatially differing models. Further, note that by inverting the property of interest, one could always swap the relative conditions of eqs (8) and (9) in order to enable a more convenient unit. For example, working with electrical resistivity $\varrho = \frac{1}{\sigma}$ instead of conductivity, freshwater intrusion into a brine-saturated host rock would cause an increase of ϱ, now calling for h_+^{Δ} (eq. 8) as the criteria for anomalously resistive cell counts.

2.5 Combining model-feature criteria: Formulation of the semblance measure

We now have two decision-making components, h^{bnd} and h^{Δ}, for classifying anomalous model variability. The first decides about the anomaly's nature in an absolute way, realized through the range $[a, b]$. The second offers a relative threshold Δ^{lim}, where both sign and magnitude of a property change with respect to a baseline determine whether this threshold is met.

The decision-making components appear as a product in the following general form for quantifying a 3D anomalous volume:

$$V^a(t) = \int_\Omega h^{bnd}(m^t_{x,y,z}, a, b) \cdot h^\Delta(\Delta^t_{x,y,z}, \Delta^{lim}) dV. \tag{10}$$

The integral aggregates non-zero volume contributions dV pertaining to feature-matching elements $m^t_{x,y,z}$ of a 3D domain Ω, where the superscript t denotes the current model state with respect to a baseline $\mathbf{m}^{t=0} = \mathbf{m}^0$ and $\Delta^t_{x,y,z}$ quantifies a time-lapse property evolution over the period t. As seen in Fig. 2, the choice of the function symbol h stems from the fact that the criteria-enforcing functions h^{bnd} and h^Δ resemble Heaviside step functions, $H(x)$.

This work employs a discrete version of eq. (10), where the domain Ω comprises a total of M grid elements. Then, the volume count becomes a cell count. Written for the reference (true) model $\tilde{\mathbf{m}}$:

$$\tilde{N}^a(t) = \sum_{i=1}^{M} h^{bnd}(\tilde{m}^t_i, a, b) \cdot h^\Delta(\tilde{\Delta}^t_i, \Delta^{lim}). \tag{11}$$

The count $\tilde{N}^a \in \mathbb{N}$ is the sum of the reference model's grid elements that meet both the boundary condition (eq. 5) and the relative condition, the latter either active as eq. (8) (h^Δ_+) or eq. (9) (h^Δ_-). In order to arrive at a similarity-measure N^a for a corresponding image $\mathbf{m}$, the conditions for the reference $\tilde{\mathbf{m}}$ need to be linked to those for $\mathbf{m}$, hence

$$N^a(t) = \sum_{i=1}^{i=M} h^{bnd}(m^t_i, a, b) \cdot h^{bnd}(\tilde{m}^t_i, a, b) \cdot h^\Delta(\Delta^t_i, \Delta^{lim}) \cdot h^\Delta(\tilde{\Delta}^t_i, \Delta^{lim}). \tag{12}$$

The link is given by the products $h^{bnd}(m^t_i, a, b) \cdot h^{bnd}(\tilde{m}^t_i, a, b)$ and $h^\Delta(\Delta^t_i, \Delta^{lim}) \cdot h^\Delta(\tilde{\Delta}^t_i, \Delta^{lim})$. Effectively, for the two models $\tilde{\mathbf{m}}$ and $\mathbf{m}$, it only lets spatially coinciding occurrences of $h^{bnd} > 0$ and $h^\Delta > 0$ pass. Otherwise, the two measures $\tilde{N}^a$ and N^a would remain independent of each other.

Finally, the two expressions $\tilde{N}^a$ and N^a combine to a ratio for quantifying the model likeness, or semblance, which may be conveniently used as a percentage,

$$S(t) = \frac{N^a(t)}{\tilde{N}^a(t)} \cdot 100. \tag{13}$$

The ideal case of a perfectly reconstructed model $\mathbf{m} = \tilde{\mathbf{m}}$ would yield $N^a = \tilde{N}^a$ and thus a semblance of S=100 %. The other extreme of S=0 can occur in a number of cases. For example, $\mathbf{m}$ might not exhibit any evolving properties that meet $h^\Delta(\Delta^t) = 1$; or, the changes actually meet $h^\Delta(\Delta^t) = 1$, however the respective model elements do not arrive at values that are considered anomalous according to the boundary condition $h^{bnd}(a, b)$; or, both criteria are met, that is, $h^{bnd}(a, b) \cdot h^\Delta(\Delta^t) = 1$, however, they do not coincide spatially with the reference, thus $h^{bnd}(m^t) \cdot h^{bnd}(\tilde{m}^t) \cdot h^\Delta(\Delta^t) \cdot h^\Delta(\tilde{\Delta}^t) = 0$. Obviously, these circumstances can occur partially so that $0 < S < 100$.

2.6 Possibilities for problem-specific function adaptation

There exists a range of possibilities for a problem-specific adaptation of the semblance concept. First, the property variations captured by the two function types h^{bnd} and h^{Δ} can become arbitrarily granular, meaning that one could define an individual pair h_i^{bnd} and h_i^{Δ} for each element i of Ω. Moreover, h_i^{Δ} might be chosen to be either h_{+i}^{Δ} (eq. 8) or h_{-i}^{Δ} (eq. 9). This would allow for the flexibility of describing multiple model features of a different nature, like electrically resistive and conductive zones. On the other hand, eq. (12) can be tailored to represent very simplistic model anomalies. The obvious choice is to assign one function pair h^{bnd} and h^{Δ} to the whole domain. Also, eq. (12) represents a more general case in order to account for both an anomalous final state, defined by h^{bnd}, and its evolution with respect to some initial state, defined by h^{Δ}. Either condition can be omitted in a simpler adaptation. Lastly, the product in eq. (10) in principal represents a series of model-feature criteria, $\Pi_j h^j$. Certain model features could inspire different or additional types of conditions, realized by functions h^j that differ from steps.

All numerical examples to be presented in the following employ discrete step functions for h^{bnd} and h^{Δ}, eqs (5)-(9). However, for the sake of completeness, I want to consider the case where it may be desirable to transition more gradually from the passing to the failing, or vice versa, of these criteria, for example when property boundaries in a reference model (or the knowledge about it) are indeed fuzzy. Adequate continuous counterparts for h^{bnd} and h^{Δ} can be constructed out of ascending and descending sigmoidal function branches. These modifications are named, respectively, t^{bnd} or t^{Δ}, because they employ the hyperbolic tangent function

$$\tanh(x) = \frac{\exp(2x) - 1}{\exp(2x) + 1} \tag{14}$$

to approximate step-like Heaviside functions. Two such approximations can then be paired up to mimic the square-function shape of the boundary condition (eq. 5),

$$t^{bnd}(m_i^t, a, b) = \begin{cases} \frac{1}{2}\big(1 + \tanh(\frac{m_i^t - a}{\epsilon})\big) & \text{for} \quad m_i^t \leq \frac{a+b}{2}, \\ 1 - \frac{1}{2}\big(1 + \tanh(\frac{m_i^t - b}{\epsilon})\big) & \text{for} \quad m_i^t > \frac{a+b}{2}, \end{cases} \tag{15}$$

where the parameter ϵ controls the steepness of t^{bnd} at the boundaries a and b; specifically, t^{bnd} approaches the step version h^{bnd} with $\epsilon \to 0$. Fig. 2a exemplifies t^{bnd} with ϵ=0.1. The functions t_+^{Δ} and t_-^{Δ} resemble step-on and step-off functions, respectively, and thus use the approximation, eq. (14), in a similar manner,

$$\begin{aligned} t_+^{\Delta}(\Delta_i^t, \Delta^{lim}) &= \quad \frac{1}{2}\Big(1 + \tanh\big(\frac{\Delta_i^t - \Delta^{lim}}{\epsilon}\big)\Big), \\ t_-^{\Delta}(\Delta_i^t, \Delta^{lim}) &= 1 - \frac{1}{2}\Big(1 + \tanh\big(\frac{\Delta_i^t - \Delta^{lim}}{\epsilon}\big)\Big). \end{aligned} \tag{16}$$

Note that another degree of freedom is given by choosing the slope-controlling ϵ-parameter different

to the one pertaining to t^{bnd}. The examples for t^{Δ}_{+} and t^{Δ}_{-} in Fig. 2b employ $\epsilon = 0.01$. A final note is that an alternative to the hyperbolic eq. (14) is given by another type of sigmoidal function, $f(x) = \frac{1}{1+\exp^{-x}}$, common in machine-learning classification problems.

3 RESULTS

The following numerical examples have the main purpose of benchmarking the proposed semblance measure against distance-based errors. The comparisons intend to highlight where the semblance can be superior in capturing preset model features – when limited model resolution renders images diffuse in comparison to a sharp reference.

3.1 Motivation: Quantifying model likeness versus absolute discrepancy

In Fig. 3, the reference model (a) is a simple circular homogeneous anomaly, to be approximated by two image instances (b and c). These images could be outcomes of a low-frequency-EM or ERT data inversion for electrical resistivity, ϱ. The objective for the semblance measure shall be to identify the resistive nature of the circular target region (Fig. 3a). Given its actual value of $\varrho = 10\ \Omega\mathrm{m}$, the semblance parameters are selected as $[a, b] = [5, \infty]$, which lets image cells $m_i = \varrho_i \geq 5\ \Omega\mathrm{m}$ contribute to the cell count N^a. Here, only the condition enforced by h^{bnd} is active in eq. (12). Note that in Fig. 3a, the tabulated errors for RMSE (eq. 1) and MAE (eq. 2) are calculated between the $1\ \Omega\mathrm{m}$ (homogeneous) background and the true model, in order to have a baseline. The percentages RMSE(%) and MAE(%) reported for Image 1 (b) and Image 2 (c) are with respect to this baseline.

Image 1 underestimates the actual anomaly's magnitude, producing roughly similar metrics for the likeness to the true model (compare RMSE(%), MAE(%), and S). The metrics appear to agree with the visual perception that about a quarter of the anomaly is reproduced in terms of its magnitude. Image 2 (Fig. 3c) overestimates the true magnitude by an order of magnitude, now leading to disagreement between the mean errors and the semblance (S). Both RMSE and MAE are above their baseline (tabulated under a), producing large negative percentages, implying a worse model fit than given by the ($1\ \Omega\mathrm{m}$) background. On the other hand, the semblance reports a high model likeness of 84 %.

The difference-based metrics perform as expected by penalizing the strong resistivity overshoot in Image 2 through large errors. However, this behavior can become undesirable given the fact that the inversion of low-frequency-EM or ERT data often leads to such imaging artifacts owing to sharp sensitivity variations near sensors or strongly contrasting property boundaries (e.g., Carey *et al.* 2017; Costall *et al.* 2018). Then, recasting the imaging objective into a more forgiving one that primarily identifies the resistive zone with its magnitude deemed secondary, can be a more useful evaluation

criteria. With this objective, the semblance measures reported for the images in Figs 3b and c appear in line with the perceived anomaly reproduction.

3.2 Example 1: Measuring the resolution loss in smooth images

The non-uniqueness of overparameterized geophysical inversions makes model smoothness a common trait, whether for one-dimensional (e.g., Constable *et al.* 1987) or multi-dimensional imaging with pixilated model spaces (e.g., Lailly & Sinoquet 1996). Smoothness can further be caused by a low degree of resolution due to low signal frequencies, as common for EM and magnetotelluric methods. This section thus attempts to shed light on the question how smoothness affects numerical model comparison.

The example in Fig. 4 is based on models of electrical conductivity from survey-design studies associated with a future fluid-injection pilot study in a coastal area near Panama City (Florida) (Jiao *et al.* 2017). With the overarching goal of testing subsurface pressure management methods for industrial-scale carbon capture and sequestration (Birkholzer *et al.* 2012), crosswell and surface-to-borehole EM surveys accompanying a planned injection schedule are considered in order to aid the spatiotemporal plume mapping.

The fluid injection scenario under investigation involves time-lapse monitoring, where the pre-injection (t=0) state will be surveyed in order to establish a baseline model $\mathbf{m}^0$, while the post-injection anomaly, representing $\tilde{\mathbf{m}}^t$, is to be mapped at some time $t > 0$. In Fig. 4, the baseline or background model (a) represents the state before injection of electrically resistive freshwater into a deep saline reservoir. Let us assume that two EM surveys are conducted, first at t=0, then after one year of injection, with the latter trying to map the evolving resistivity anomaly in Fig. 4b. This anomaly is the result of a 3D fluid flow- and transport simulation using the TOUGH2 flow and transport simulator (Pruess 2004) and a subsequent petrophysical transformation from brine saturation into electrical conductivity. Underlying porosity and hydraulic permeability depth profiles are based on a layered geological model of the Upper Cretaceous sandstones belonging to the lower Tuscaloosa Formation (González-Nicolás *et al.* 2019). A peculiarity in the geological model is the low-permeability layer around the depth of 1490 m, causing a split plume, which would pose a challenge to any smoothness-constrained inversion method. Further hydrogeophysical details are omitted here for brevity, because rather than performing actual inversions, smooth images are mimicked here by applying a spatial (7-point Laplacian) smoothing filter to the true model $\tilde{\mathbf{m}}^t$ (Fig. 4b). Fig. 4 exemplifies three increasing degrees of image smoothness, realized by 5 (c), 50 (d), and 100 (d) filter sweeps.

It is now of interest to look at the trend that the different fit qualifiers show with increasing image blurriness. Therefore, the smoothed versions of $\tilde{\mathbf{m}}^t$ now become the input $\mathbf{m}^t$, representing hypothet-

ical inversion outcomes. Fig. 5a compares their semblance S, calculated using eqs. (11)-(13), against the difference terms ε_{RMSE} (eq. 1) and ε_{MAE} (eq. 2). Fig. 5b repeats the comparison for the logarithmic variants ε^{log}_{RMSE} (eq. 3) and ε^{log}_{MAE} (eq. 4). For the sake of a clearer comparison, S is first recast into the variant

$$S' = 1 - \frac{S}{100}, \;\; [S] = \%, \tag{17}$$

so that the minimum (S=0 %) and maximum (S=100 %) semblances translate to the respective errors S'=1 and S'=0. Given the goal of pinpointing a resistive anomaly evolving within a conductive background, the latter averaging $\sigma^0 \approx 3$ Sm^{-1}, the semblance parameters are set to $[a, b]$=$[0, 1.1]$ Sm^{-1} and $\Delta^{lim} = -0.2$ (eq. 7). In other words, only imaging grid elements i of $\mathbf{m}^t$ characterized by $\sigma_i \leq 1.1$ Sm^{-1} and a conductivity decrease by at least 20 % ($\Delta_i \leq \Delta^{lim} = -0.2$) with respect to the background produce non-zero counts in eq. (12).

Errors to the left of the x-axis in Fig. 5 marked as "true" are "auto-errors", since their input is $\mathbf{m}^t = \tilde{\mathbf{m}}^t$ (the true model, Fig. 4b), producing the consistency check $\varepsilon = S' = 0$ for all candidates. Errors at the left end of the x-axis (marked as "background") correspond to the baseline (Fig. 4a), that is, $\mathbf{m}^t = \mathbf{m}^0$ in all error terms. In inversions of post-injection time-lapse data sets, $\mathbf{m}^0$ would most likely serve as the starting model; hence corresponding errors are expected to be maximal, which holds for ε_{RMSE}, ε^{log}_{RMSE}, and S' (note that S'=1 exceeds the shown plot range). In contrast, this does not hold for ε_{MAE} and ε^{log}_{MAE} as they already surpass the background value after a few smoothing sweeps. These different trends indicate that absolute differences terms are more susceptible to overrating error contributions due to smoothed boundaries than squared differences.

In conclusion, this example shows that for time-lapse imaging cases with ample a priori information about evolving anomalies, here provided by the underlying flow modeling, the semblance measure appears most robust in terms of suppressing unwanted error contributions due to image smoothness. Assuming that each smoothing sweep deteriorates the image quality by an equal amount, an ideal likeness metric would show a linear error increase. Two observations in Fig. 5 indicate that the semblance comes closest to that. First, while all errors increase the fastest during the first smoothing sweeps, this increase is the slowest for S'. This is desirable because a moderately smooth image as exemplified by Fig. 4c (after 5 smoothing sweeps) would still be considered of good quality – at least in EM imaging contexts – given that the major features of the split plume remain visible. Second, with a large number of smoothing sweeps, errors converge towards a value near (but not necessarily leading to) the baseline value; this convergence is now the fastest for S', which is also desirable when the goal is to distinguish between subtle image differences at the low-quality end.

3.3 Example 2: Evaluating permeability images of hydrogeophysical inversions

This example intends to demonstrate the semblance formulation for imaging problems where a baseline state $\mathbf{m}^0$ is not involved, that is, a time-lapse property evolution is unknown or not of interest. For this purpose, eqs (11) and (12) are modified by simply omitting all relative-condition terms h^{Δ}.

This and the following example's goals are to appraise the image quality of (hydraulic) permeability fields and their resulting forecasting of tracer flow for a simulated near-surface fluid injection experiment recently reported in a didactic contribution (Commer *et al.* 2020). Specifically, how well does visually perceived image quality shown by Fig. 6 reflect in all error options?

In subsurface flow predictions, permeability is paramount as it is a major hydrogeological property controlling preferential fluid flow paths. It is often abbreviated as k and uses the SI unit m^2 (another common unit is Darcy, where 1 Darcy $\approx 10^{-12}\,\mathrm{m}^2$). Owing to its typically large range over several decades, the following error calculations employ a logarithmic (base 10) transform.

The inverse solutions to be evaluated originate from a hydrogeophysical data set simulated for a fluid injection into a shallow aquifer. The underlying true permeability field is shown in Fig. 6a together with the synthetic-data experimental setup. Injection of an electrically conductive (brine-enriched) tracer occurs over the screened interval at the left end. Flow along the x-axis is driven by a constant pressure gradient. Figs 6b, c, and d are permeability images obtained from three hydrogeophysical synthetic-data inversions with the aim of replicating the actual (a) permeability distribution. Fig. 6b results from inverting in-situ (extracted over an array of four monitoring wells) tracer concentration measurements (referred to as C-data). Fig. 6c results from inverting crosswell electrical resistivity tomography (ERT) data sampled over an array of four instrumented wells. Fig. 6d involves the joint inversion of both data types. Each inversion employed a homogenous field of $k = 10^{-11}\,\mathrm{m}^2$ as starting model. For a discussion of the image differences from a hydrogeophysical perspective, I again refer to Commer *et al.* (2020) as, for brevity, only relevant details are provided here.

Each model-vector pair ($\tilde{\mathbf{m}}$,$\mathbf{m}$) (where the time superscript is not needed) entering the error calculations is represented by the true permeability model in Fig. 6a and one of the three images, respectively. Further, proper semblance parameters need to be predefined. Environmental or reservoir-monitoring fluid-flow simulations often are predominantly interested in mapping preferential flow paths. This can be realized by setting $[a, b]$=$[-10.5, \tilde{k}_{max}]$, where $\tilde{k}_{max}$ is the true model's maximal permeability (or any value above), thus selecting all image elements with a permeability $\log_{10}(k) \geq -10.5$, or $k \geq 3.16 \times 10^{-11}\,\mathrm{m}^2$. Discerning between both low-k and high-k regions, a second semblance realization will be used, to be called S_{l+h} as it uses two boundary pairs $[a, b]_l$=$[\tilde{k}_{min}, -11.5]$ and $[a, b]_h$=$[-10.5, \tilde{k}_{max}]$.

Table 1 lists the suite of errors for each permeability image. Errors involve $\log_{10}(k)$-terms, hence

the log error variants (eqs 3 and 4) are omitted. In order to highlight the error improvement (or lack thereof) of the ε-terms, errors are first calculated with the starting model as input $\mathbf{m}$ (annotated as "Starting model"). In addition to the inversion result (annotated as "Final"), the percentage change with respect to the starting model is calculated (annotated as "%").

Negative percentages for the ε-terms indicate a poor or failed model reproduction. The hydrological data inversion (using only C-data) exhibits relatively large negative percentages, which is in agreement with the low values for both instances of S. Inverting only ERT-data shows the largest error decrease for ε_{RMSE} and ε_{MAE}, while the joint inversion yields poor ratings. The reason can be seen in the joint inversion image (Fig. 6d) exhibiting minor image artifacts. Owing to the large permeability range, artifacts can disproportionately elevate the differences in ε_{RMSE} and ε_{MAE}. Recall that the semblance formulation does not overrate differences between $\tilde{\mathbf{m}}$ and $\mathbf{m}$. Hence, it is plausible that the joint inversion produces the highest semblance, because the permeability magnitude in the low-k and high-k zones is replicated better than by the ERT-data inversion.

3.4 Example 3: Assessing the forecasting ability of a hydrogeophysical inversion

Fig. 6 reveals a discretization difference between the grid used for the hydrological forward modeling (a), where $\tilde{\mathbf{m}}$ (and Ω) resides, and the imaging grid (b, c, and d). The latter was made coarser in order to reduce the number of unknowns comprising $\mathbf{m}$. To account for this discretization disparity, all error calculations are preceded by a mapping of $\mathbf{m}$ onto the fine grid.

For the degree of resolution that the permeability model shows in Fig. 6a with respect to the exemplified field scale, it could be shown that there exists an averaging effect along contrasting facies on preferential paths of tracer advection (Commer *et al.* 2020). In other words, while the coarser imaging grid introduces a certain resolution loss in the permeability images, as also shown by the generally poor ratings in Table 1, this loss carries over only marginally to subsequent flow forecasting. Visually, this reveals in Fig. 7 in a good prediction of the tracer concentration by those permeability models that approximate the major flow paths (Fig. 6c and d).

Fig. 7 presents snapshots of the tracer flow over a 161-day period, selecting the flow times t=30, t=70, t=110, and t=161 days. Each plot row (a-d) corresponds to a hydrological forward-modeling run with its respective permeability model shown by Figs 6a-d. Images originating from post-inversion predictive forward-modeling of this kind differ from the oftentimes fuzzy output of underdetermined inversions in that contrasts of model attributes are purely caused by the physics posed by the fluid-injection process, rather than by a regularized inversion procedure. This leads to a similar degree of image sharpness between the true case (Fig. 7a) and its predictions (Figs 7b-d).

Table 2 presents error calculations for all 12 images of Fig. 7b-d. Note that the true model (Fig. 7a)

$\tilde{\mathbf{m}}$ is now a vector of tracer concentration mass fractions $\tilde{\mathbf{C}}$ ranging from zero to one, whereas the images represent the input $\mathbf{m} = \mathbf{C}$. With the background (pre-injection) state being C=0 over Ω, the logarithmic errors are omitted. Semblance parameters are $[a, b] = [0.75, 1]$ chosen to identify all image pixels i with a tracer content $C_i \geq 0.75$. No relative condition is enforced.

The similar degree of image sharpness leads to a generally closer alignment between the trends indicated by ε_{RMSE}, ε_{MAE}, and S. On average, all errors signify that the permeability model stemming from the ERT-data inversion (Fig. 6c) achieves the closest replication of the actual tracer flow. This appears to support the visual impression that the joint inversion's minor permability image artifacts lead to a slightly deteriorated tracer-field prediction (compare Fig. 7c against Fig. 7d).

4 CONCLUSIONS

The proposed goodness of fit measure, called semblance S, is designed for model inter-comparison in application scenarios with a rather qualitative objective that identifies the nature of model features, as opposed to quantitative global fitting of model parameters or their data predictions. The semblance metric provides an alternative to distance-based metrics and may be useful for the numerical rating of potentially numerous imaging outcomes with equivalent data fits. Machine-learning applications for survey design where synthetic-data inverse modeling steers the decision-making of an overarching optimization scheme may be an upcoming application area.

Instead of traditionally used least-squares error types, the semblance is based on the idea of logistic functions. Logistic function types model a binary dependent variable, which in this case has the output $f = h^{bnd}h^{\Delta}$ with f=1 for passing or f=0 for failing preset boundary and relative-change anomaly criteria for a given target. This function behaviour makes the semblance more forgiving towards imaging over- or undershoots than distance-based metrics. Again, this may be desirable if the quality of a model feature is of primary interest, as opposed to exactly matched quantities.

The numerical examples demonstrated that the semblance avoids spurious contributions that occur for difference measures when a reference model's well-defined anomaly contours become smoothed in its image. RMSE and MAE error calculations from images without smoothed contrast boundaries exhibited a clearer correlation to the semblance ratings, indicating a similar model evaluation performance. All example images demonstrated a good agreement between visually perceived image quality in terms of likeness to its reference and corresponding S ratings. I thus regard the semblance metric as an additional flexible tool, in some way comparable to a zoom lens, within the portfolio of metrics with different characteristics for image analysis. While it may complement, it can by no means replace any metrics for statistical analysis.

Selection of proper semblance parameters is straightforward for synthetic-data inversions, where the target is known. The combination of absolute and relative conditions, or the omission of either one, offers flexibility in predefining target features. For survey-design purposes, this target-adjusting capacity of S may be beneficial in rating trial inversions that are distinguished by only subtle image differences.

While not pursued in this contribution, every functional relationship for model inter-comparison can in principal be cast into a model-constraining objective functional driving an inversion. The demonstrated robustness in discerning anomalies in smooth images may make the semblance a valid alternative to the cross-gradient method (Gallardo & Meju 2003). Cross-gradients are often used in joint inversions as a model-constraining objective functional that enforces structural model similarity. However, there is evidence that model features that are smooth with respect to a sharper reference can compromise the cross-gradient linkage as a minimizer due to spurious contributions along actually aligning boundaries (Maysami & Clapp 2009). As an alternative, and given that smooth models are ubiquitous in geophysical imaging, future research may try the sigmoidal-function implementation (eqs 14-16) of S for model linkage in joint inversions.

ACKNOWLEDGMENTS

The theoretical part of this work was supported by the United States Department of Energy, Office of Science, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under Contract DE-AC02-05CH11231. One presented example was designed for field-design purposes supported by the U.S. DOE under Award Number DE-FE0026140, "Gulf Coast Field Demonstration at a Flagship Power Plant to Assess Optimal Reservoir Pressure Control, Plume Management and Produced Water Strategies". This paper was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

I am grateful to Colin G. Farquharson, an anonymous reviewer, and editor Andrea Morelli for very constructive feedback. Last but not least, I dedicate this paper to my friend and mentor Gregory A. Newman in honor of his career and upon his recent retirement.

REFERENCES

Birkholzer, J.T., Cihan, A. & Zhou, Q., 2012. Impact-driven pressure management via targeted brine extraction? Conceptual studies of CO2 storage in saline formations, *International Journal of Greenhouse Gas Control*, **7**, 168-180.

Botchkarev A., 2019. A new typology design of performance metrics to measure errors in machine learning regression algorithms, *Interdisciplinary Journal of Information, Knowledge & Management*, **14**, 45-76.

Carey, A., Paige, G., Carr, B. & Dogan, M., 2017. Forward modeling to investigate inversion artifacts resulting from time-lapse electrical resistivity tomography during rainfall simulations, *J. Appl. Geophys.*, **145**, doi:10.1016/j.jappgeo.2017.08.002.

Caruana, R., 1993. Multitask learning: A knowledge-based source of inductive bias, in *Proceedings of the Tenth International Conference on Machine Learning*, 41-48.

Chai, T., & Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature, *Geoscientific Model Development*, **7**, 1247-1250.

Commer M., Pride S.R., Vasco D.W., Finsterle S. & Kowalsky M.B., 2020. Imaging of a fluid injection process using geophysical data - a didactic example, *Geophysics*, **85**, W1-W16.

Constable, S.C., Parker, R.L. & Constable, C.G., 1987. Occam's inversion: A practical algorithm for generating smooth models from electromagnetic sounding data, *Geophysics*, **52**, 289-300.

Costall, A., Harris, B. & Pigois, J.P., 2018. Electrical resistivity imaging and the saline water interface in high-quality coastal aquifers, *Surv. Geophys.*, **39**, 753-816.

Deza, M. & Deza, E., 2009. *Encyclopdedia of distances*, Springer, Dordrecht Heidelberg.

Farquharson, C.G., 2008. Constructing piecewise-constant models in multidimensional minimum-structure inversions, *Geophysics*, **73**, K1-K9.

Gallardo, L.A. & Meju, M.A., 2003. Characterization of heterogeneous near-surface materials by joint 2D inversion of dc resistivity and seismic ata. *Geophys. Res. Lett.*, **30**, 1658.

González-Nicolás, A., Cihan, A., Petrusak, R., Zhou, Q., Trautz, R., Riestenberg, D., Godec, M. & Birkholzer, J.T., 2019. Pressure management via brine extraction in geological CO2 storage: adaptive optimization strategies under poorly characterized reservoir conditions, *Int. J. Greenhouse Gas Control*, **83**, 176-185.

Huber-Carol, C., Balakrishnan, N., Nikulin, M.S. & Mesbah, M., (eds.), 2002. *Goodness-of-fit tests and model validity*, Birkhäuser, Boston.

Hughson, D.L. & Yeh T.-C. J., 2000. An inverse model for three-dimensional flow in variably saturated porous media, *Wat. Resour. Res.*, **36**, 829-839.

Jackson, D.D., 1979. The use of a priori data to resolve non-uniqueness in linear inversion, *Geophys. J Int.*, **57**, 137-157.

Jiao, Z., Pawar, R., Duguid, A., Bourcier, W., Haussmann, C., Coddington, K., Harp, D., Ganshin, Y., Quillinan, S., McLaughlin, F. & Ramsey, R., 2017. A field demonstration of an active reservoir pressure management through fluid injection and displaced fluid extraction at the Rock Springs Uplift a priority geologic CO2 storage site for Wyoming, *Energy Procedia*, **114**, 2799-2811.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A. & Kumar, V., 2019. Machine Learning for the geosciences: challenges and opportunities, *IEEE Transactions on Knowledge and Data Engineering*, **31**, 1544-1554.

Kellman, M., Zhang, K., Tamir, J., Bostan, E., Lustig, M. & Waller, L., 2020. Memory-efficient learning for large-scale computational imaging, *Arxiv, doi: arXiv:2003.05551*.

Kitanidis, P.K., 2002. *Introduction to geostatistics: applications in hydrogeology*, Cambridge University Press, Cambridge.

Lailly, P. & Sinoquet, D., 1996. Smooth velocity models in reflection tomography for imaging complex geological structures, *Geophys. J Int.*, **124**, 349-362.

Kowalsky, M.B., Finsterle, S., Williams, K.H., Murray, C., Commer, M., Newcomer, D., Englert, A., Steefel, C.I. & Hubbard, S.S., 2012. On parameterization of the inverse problem for estimating aquifer properties using tracer data, *Water Resour. Res.*, **48**, W06535.

Menke, W., 2018. *Geophysical data analysis: Discrete inverse theory*, 4th edn, Academic Press Inc., New York.

Maysami, M. & Clapp, R.G., 2009. The cross-gradient function: a structural similarity measure, http://sepwww.stanford.edu/data/media/public/docs/sep138/mohammad1/paper_html/node2.html, accessed 3 March 2020.

Maurer, H., Curtis, A. & Boerner, D.E., 2010. Recent advances in optimized geophysical survey design, *Gephysics*, **75**, 75A177-75A194.

Nakayama, S., Blacquière, G. & Ishiyama, T., 2019. Automated survey design for blended acquisition with irregular spatial sampling via the integration of a metaheuristic and deep learning, *Gephysics*, **84**, P47-P60.

Portniaguine, O. & Zhdanov, M.S., 1999. Focusing geophysical inversion images, *Geophysics*, **64**, 874-887.

Pruess, K., 2004. The TOUGH codes - A family of simulation tools for multiphase flow and transport processes in permeable media, *Vadose Zone J.*, **3**, 738-746.

Putzky P. & Welling, M., 2019. Invert to learn to invert, *NeurIPS, doi: arXiv:1911.10914*.

Russell, B., 2013. The S function and its geophysical applications, *CREWES Research Report*, Volume 25.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J. & Rowe, C. M., 1985. Statistics for the evaluation and comparison of models, *J. Geophys. Res.*, **90**(C5), 8995? 9005.

Willmott, C.J. & Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Climate Research*, **30**, 79-82.

Yeh, T.-C. J.& Liu, S., 2000. Hydraulic tomography: Development of a new aquifer test method, *Wat. Resour. Res.*, **36**, 2095-2105.

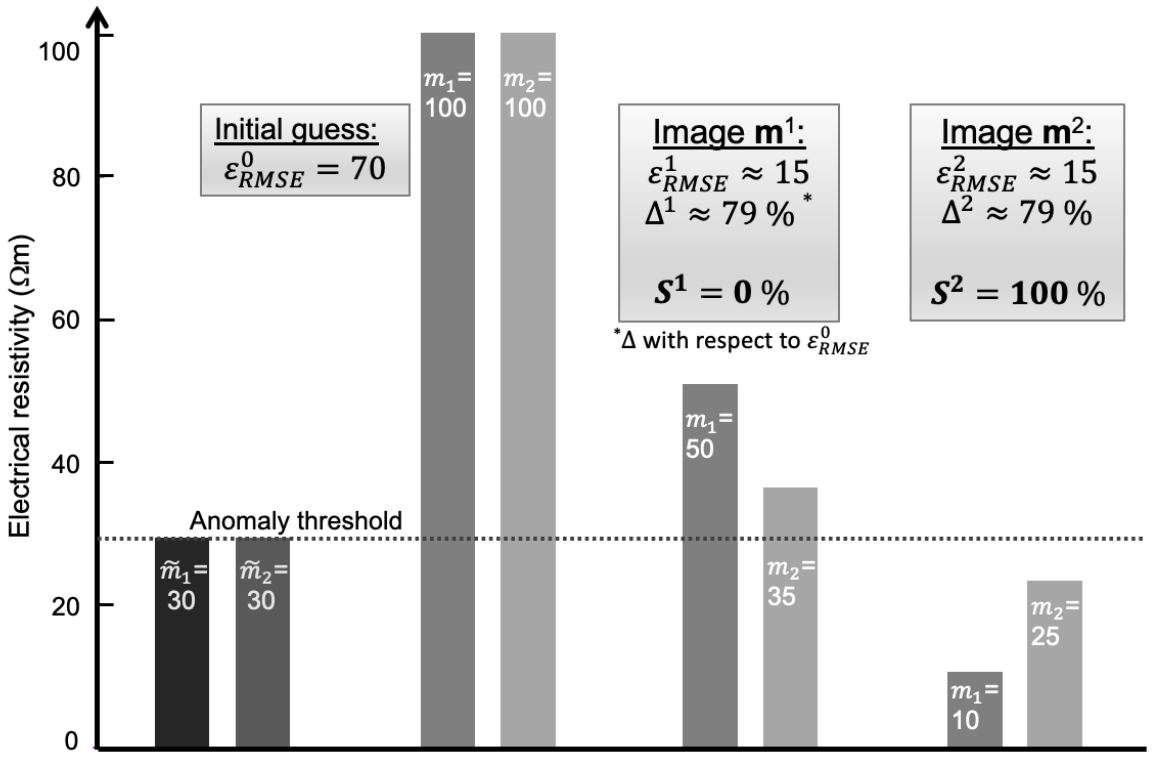

Figure 1. Illustration of a model example with two electrical resistivity parameters. The RMSE involves the term $m_i - \tilde{m}_i$, where $\tilde{m}_1 = \tilde{m}_2 = 30\ \Omega\mathrm{m}$ is the true model. Errors are calculated for a hypothetical initial guess ($\tilde{m}_1 = \tilde{m}_2 = 100\ \Omega\mathrm{m}$) and two image instances. Relative percentages Δ^1 and Δ^2 quantify the error reduction with respect to ε^0_{RMSE}. Semblance percentages S^1 and S^2 quantify how many parameters meet the criteria $m_i \leq 30\ \Omega\mathrm{m}$.

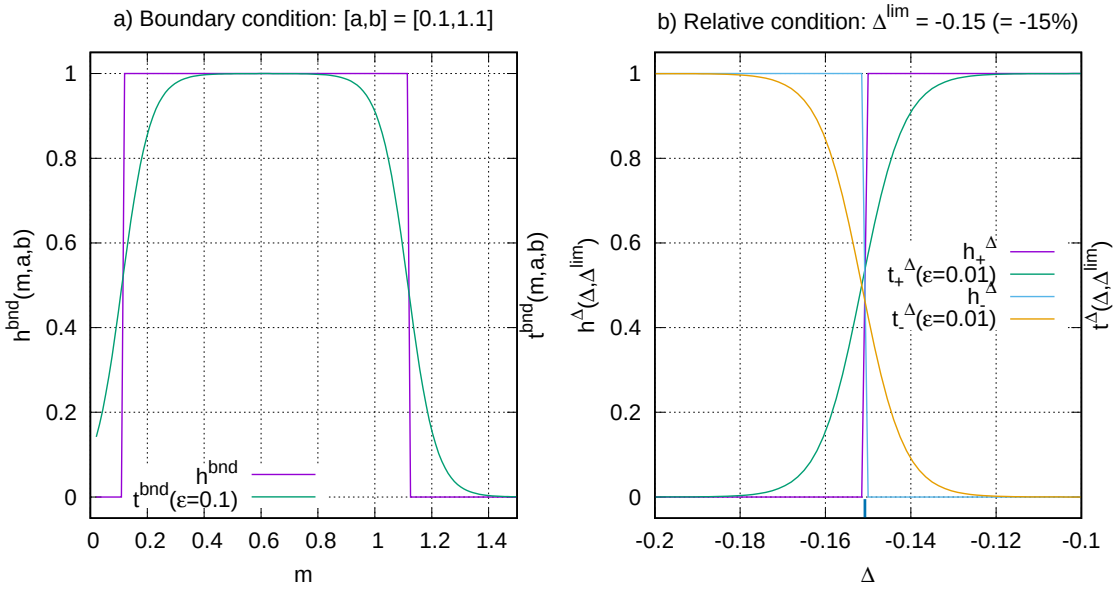

Figure 2. Function example for enforcing the boundary condition (a) and the relative condition (b). The conditions can be implemented either through step functions (symbol h) or sigmoidal function branches (symbol t).

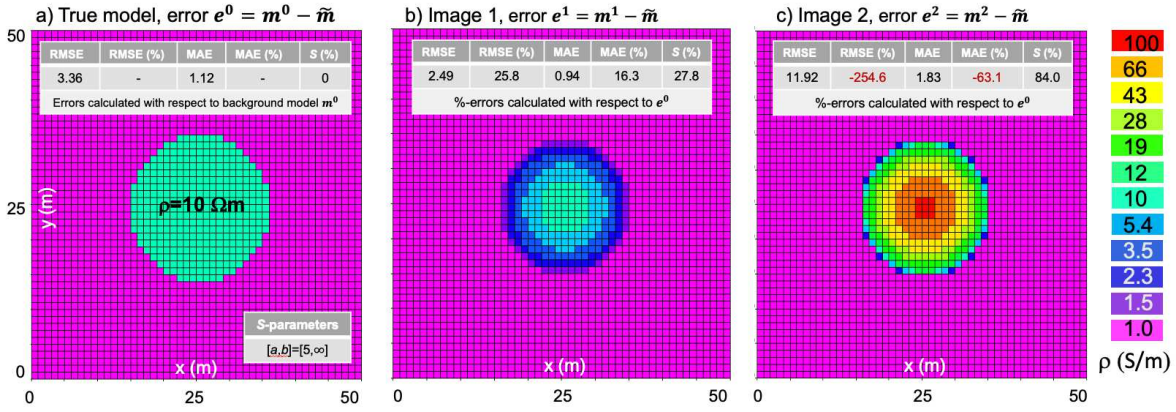

Figure 3. Example model of electrical resistivity representing a case where the semblance can be superior over difference metrics. The true model (a) is approximated by an image with underestimated resistivity magnitude (Image 1, b), and an image with strongly overestimated resistivity magnitude (Image 2, c). Relative percentages for RMSE and MAE are with respect to the baseline errors RMSE=3.36 and MAE=1.12 (tabulated under a) calculated between the true model and the homogeneous (1 Ωm) background.

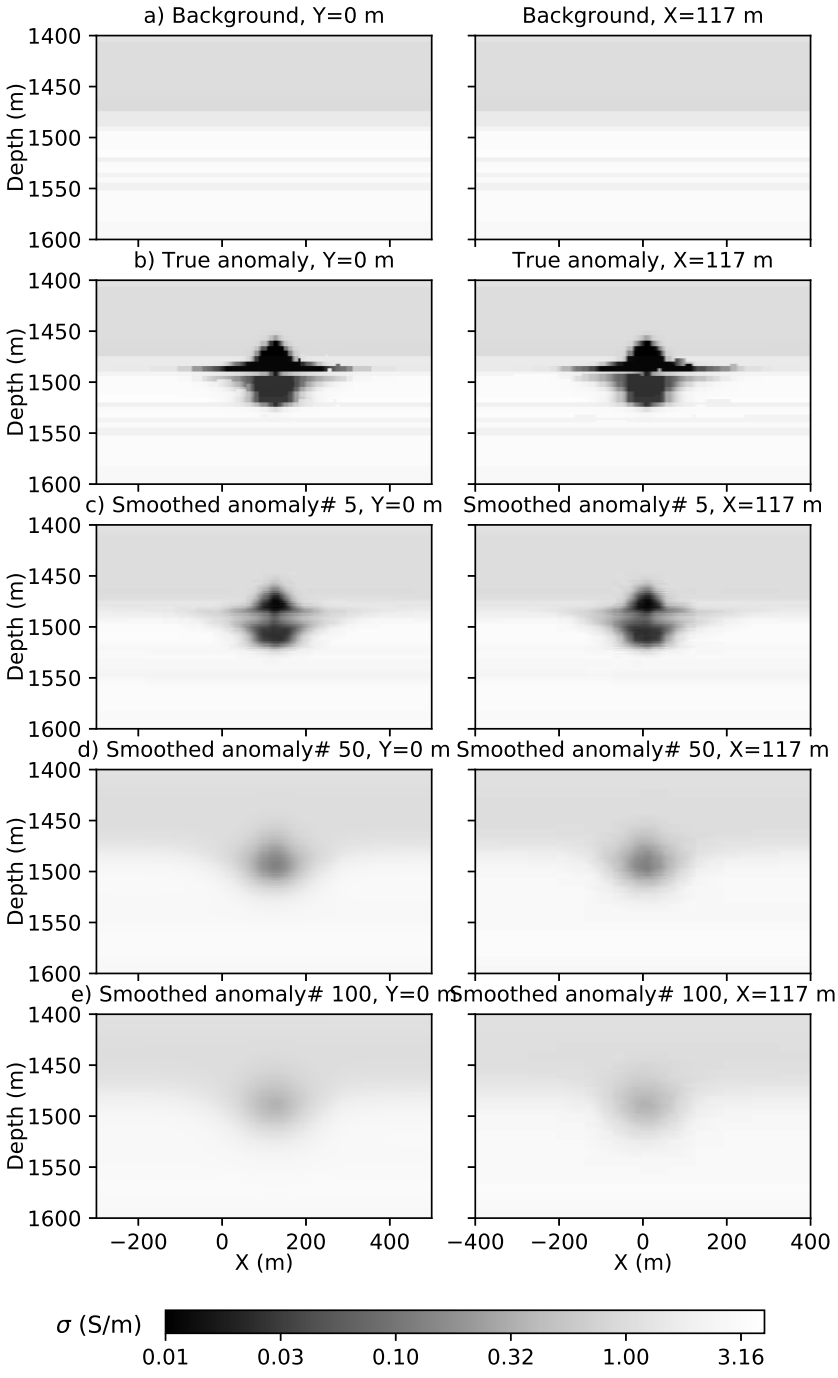

Figure 4. Background (a), true model (b), and images with increasing degrees of model smoothness (c-e).

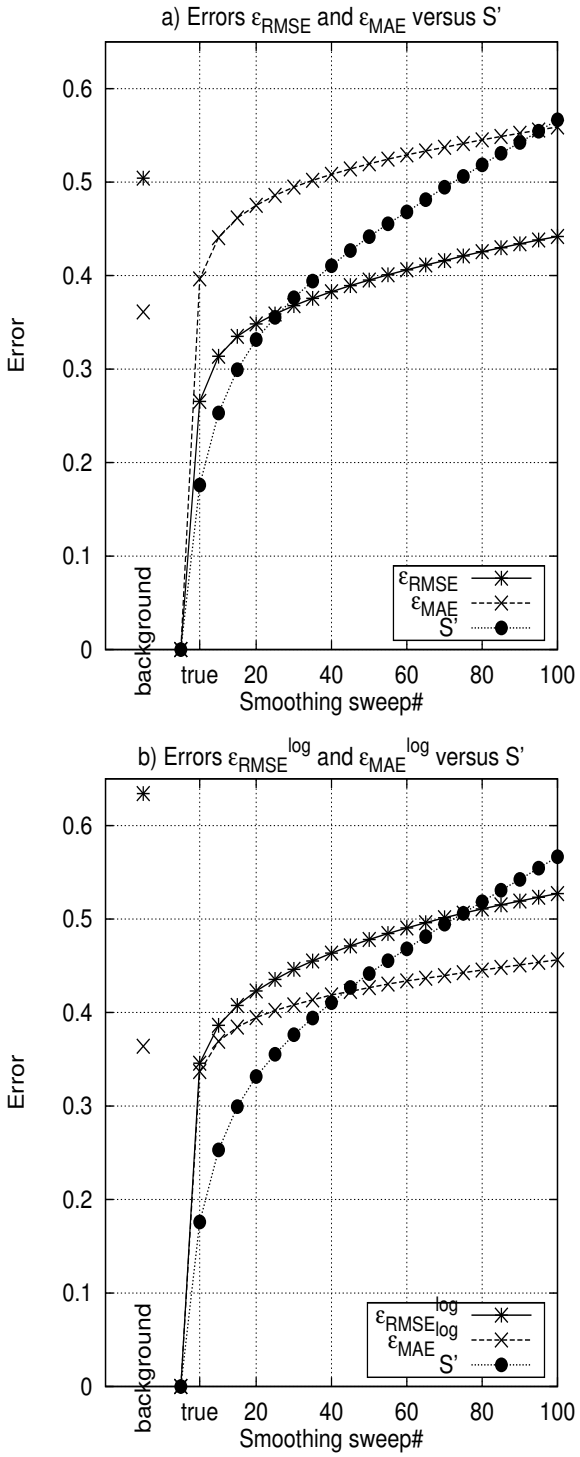

Figure 5. Error trends calculated from the smooth images of Fig. 4. Error terms ε_{RMSE} and ε_{MAE} (a) and logarithmic variants ε^{log}_{RMSE} and ε^{log}_{MAE} (b) are compared against an error equivalent S' based on the semblance S.

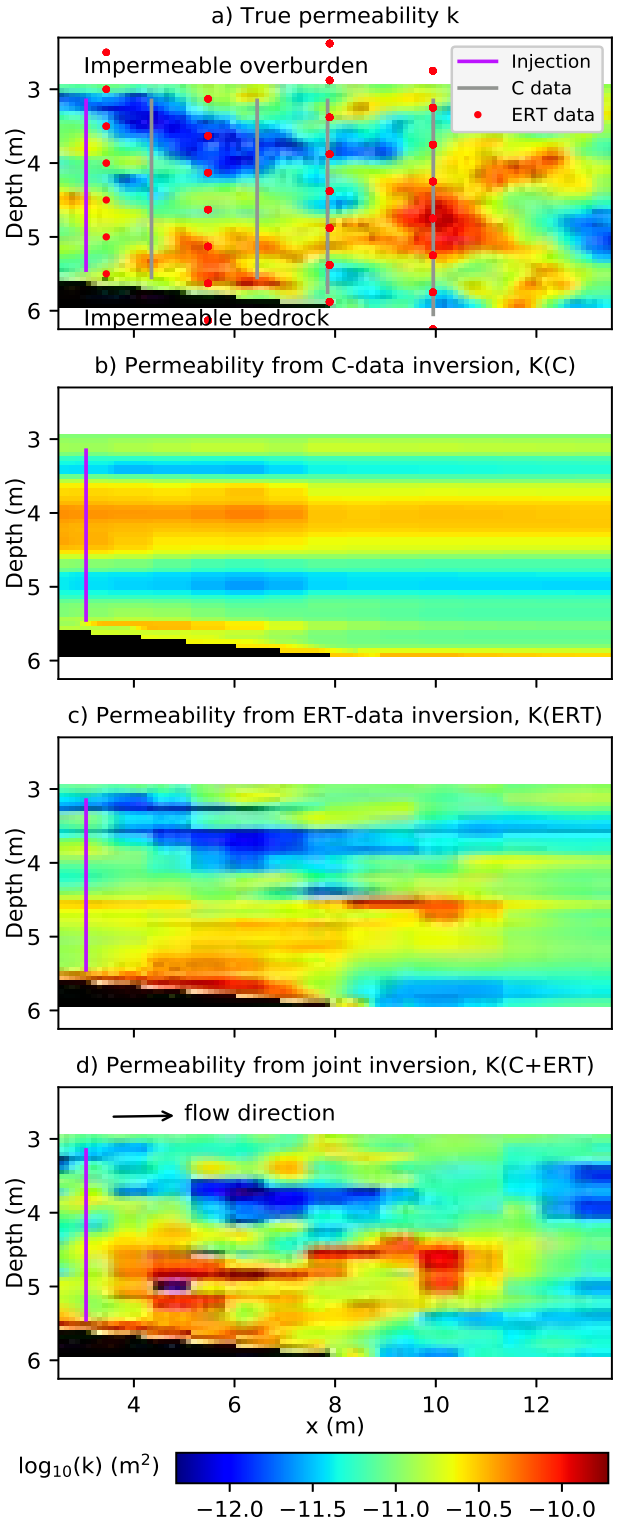

Figure 6. Actual 2D permeability distribution (a) and images resulting from inversions of hydrological (b), ERT (c), and combined (d) data sets (Commer *et al.*, 2020).

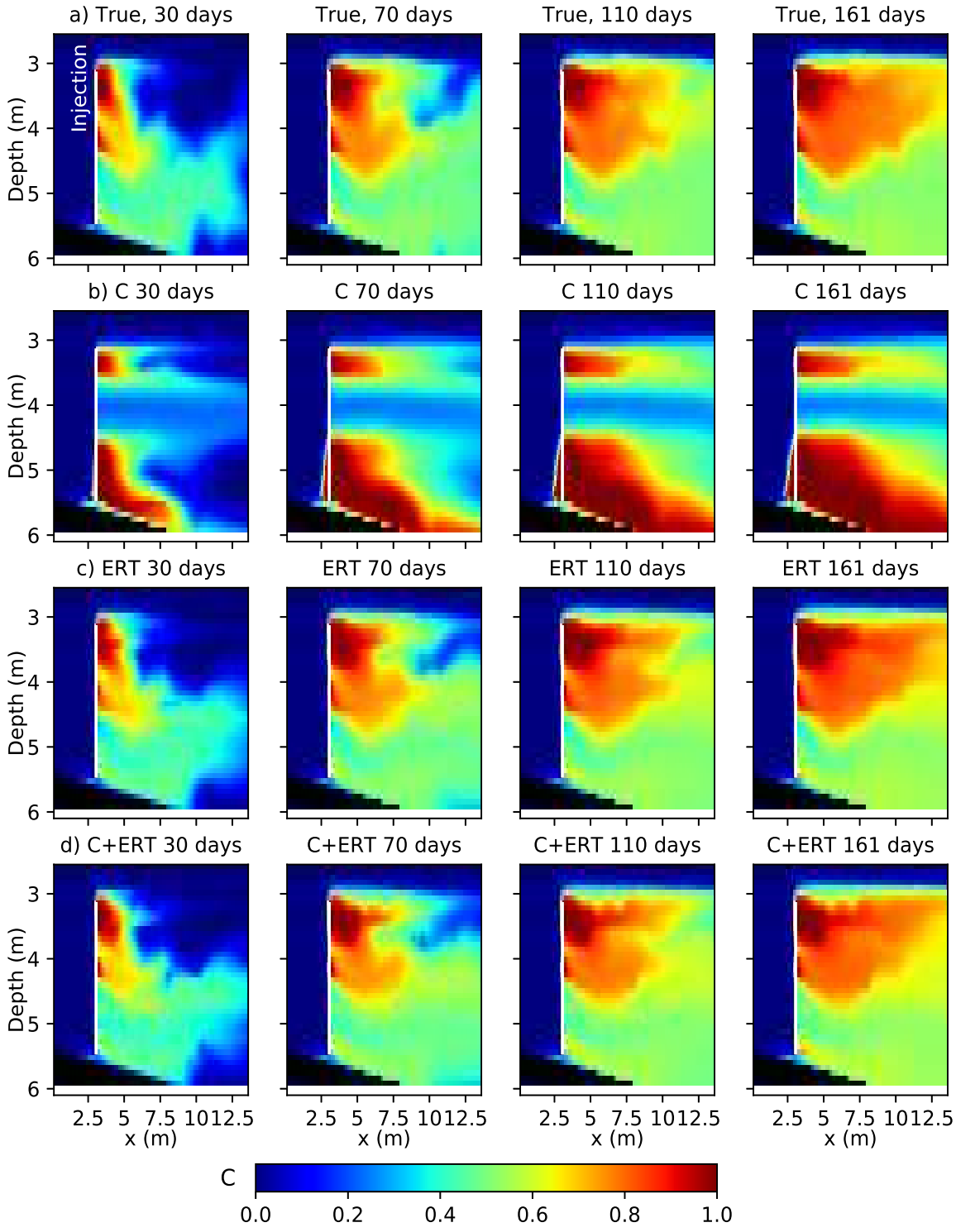

Figure 7. Forecasting of tracer flow based on the permeability predictions in Fig. 6. The true flow field (a) results from the true permeability distribution in Fig. 6a. Predicted flow fields (b-d) result from their respective images (Fig. 6b-d).

Image		ε_{RMSE}	ε_{MAE}	$S_h(\%)$	$S_{l+h}(\%)$
Starting model		0.4102	0.5391	0.0000	0.0000
C-data	Final	0.5547	0.6216	10.2669	6.8385
inversion	%	-35.20	-15.31	10.27	6.84
ERT-data	Final	0.3317	0.4762	22.7926	35.3816
inversion	%	19.13	11.66	22.79	35.38
Joint	Final	0.4241	0.5383	43.1212	43.3102
(C+ERT)	%	-3.37	0.15	43.12	43.31

Table 1. Error values for three permeability images obtained from a hydrogeophysical inversion study (Commer *et al.*, 2020). Images stem from the inversion of C-data (Fig. 6b), ERT-data (Fig. 6c), and their combination (Fig. 6d). The two semblance instances aim for high-k regions (S_h) and a combination of low-k and high-k regions (S_{l+h}), where low-k and high-k boundary parameters are $[a, b]_l$=$[\tilde{k}_{min}, -11.5]$ and $[a, b]_h$=$[-10.5, \tilde{k}_{max}]$, respectively. Note that errors calculated from the starting model are identical for all inversions.

Image	Model	ε_{RMSE}	ε_{MAE}	$S(\%)$
C-30	Initial	0.3697	0.5207	0.0000
	Final	0.2494	0.4246	33.3333
	%	32.54	18.45	33.33
C-70	Initial	0.4979	0.6445	0.0000
	Final	0.2962	0.4593	26.8354
	%	40.50	28.73	26.84
C-110	Initial	0.5542	0.6871	0.0000
	Final	0.3238	0.4910	26.5337
	%	41.58	28.54	26.53
C-161	Initial	0.5797	0.7044	0.0000
	Final	0.3415	0.5147	20.5947
	%	41.09	26.93	20.59
ERT-30	Initial	0.3697	0.5207	0.0000
	Final	0.0433	0.1549	88.4058
	%	88.28	70.25	88.41
ERT-70	Initial	0.4979	0.6445	0.0000
	Final	0.0379	0.1431	94.9367
	%	92.38	77.79	94.94
ERT-110	Initial	0.5542	0.6871	0.0000
	Final	0.0320	0.1347	95.7055
	%	94.22	80.39	95.71
ERT-161	Initial	0.5797	0.7044	0.0000
	Final	0.0336	0.1422	99.4493
	%	94.20	79.81	99.45
C+ERT-30	Initial	0.3697	0.5207	0.0000
	Final	0.0579	0.1900	81.1594
	%	84.34	63.50	81.16
C+ERT-70	Initial	0.4979	0.6445	0.0000
	Final	0.0445	0.1621	78.2278
	%	91.07	74.85	78.23
C+ERT-110	Initial	0.5542	0.6871	0.0000
	Final	0.0361	0.1472	85.8896

	%	93.49	78.58	85.89
C+ERT-161	Initial	0.5797	0.7044	0.0000
	Final	0.0363	0.1535	89.8678
	%	93.74	78.21	89.87

Table 2: Error comparisons for the example of forecasting tracer concentration (Fig. 7). Errors are calculated for the images produced by each inversion's initial and final model guess. Percentage error improvements are also calculated with respect to the initial fit.