Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR037380

Learning Distributed Parameters of Land Surface Hydrologic Models Using a Generative Adversarial Network

Ruochen Sun[1,2,3], **Baoxiang Pan**[4], **and Qingyun Duan**[1,2,3]

[1]The National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China, [2]College of Hydrology and Water Resources, Hohai University, Nanjing, China, [3]China Meteorological Administration Hydro-Meteorology Key Laboratory, Hohai University, Nanjing, China, [4]Institute of Atmospheric Physics, Chinese Academy of Science, Beijing, China

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Key Points:

- A novel generative adversarial network-based parameter estimation method is proposed to calibrate distributed land surface hydrologic models
- By employing a discriminator to identify model spatial biases, this method contributes to effective and spatially coherent parameter estimation
- This method can substantially reduce model simulated errors at grid scale and achieve consistent spatial performance

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Q. Duan,
qyduan@hhu.edu.cn

Citation:

Sun, R., Pan, B., & Duan, Q. (2024). Learning distributed parameters of land surface hydrologic models using a Generative Adversarial Network. *Water Resources Research*, *60*, e2024WR037380. https://doi.org/10.1029/2024WR037380

Received 22 FEB 2024
Accepted 27 JUN 2024

Author Contributions:

Conceptualization: Ruochen Sun, Baoxiang Pan
Formal analysis: Ruochen Sun
Funding acquisition: Ruochen Sun, Baoxiang Pan, Qingyun Duan

Abstract Land surface hydrologic models adeptly capture crucial terrestrial processes with a high level of spatial detail. Typically, these models incorporate numerous uncertain, spatially varying parameters, the specification of which can profoundly impact the simulation capabilities. There is a longstanding tradition wherein parameter calibration has served as the conventional procedure to enhance model performance. However, calibrating distributed land surface hydrologic models presents a great challenge, often resulting in uneven spatial performance due to the compression of information inherent in model outputs and observations into a single-value objective function. To address this problem, we propose a novel Generative Adversarial Network-based Parameter Optimization (GAN-PO) method. By leveraging a deep neural network to discern model spatial biases, we train a generative network to produce spatially coherent parameter fields, minimizing distinctions between simulations and observations. By leveraging neural network-based surrogate models to make the physical model differentiable, we employ GAN-PO to calibrate the Variable Infiltration Capacity (VIC) model against evapotranspiration (ET) over China's Huaihe basin. The results show that GAN-PO can diminish errors in simulated ET derived from default parameters across nearly all grid cells within the study region, surpassing the conventional calibration approach based on the parameter regionalization technique. Ablation analysis indicates that relying solely on the traditional loss could lead to deteriorated model performance, underscoring the crucial role of the discriminator. Notably, due to the discriminator's explicit identification of model spatial biases, GAN-PO excels in maintaining spatial consistency, outperforming the state-of-the-art differentiable parameter learning (dPL) method in terms of model spatial performance.

1. Introduction

Modern land surface hydrologic models can intricately represent the key terrestrial processes in a spatially detailed manner. These models are vital for understanding water and energy balances, biogeochemical cycling, as well as the interaction with weather, climate, and various earth environments (Blyth et al., 2021). A prevalent and enduring issue with the land surface hydrologic models is that their performance and capabilities are significantly influenced by spatially varying parameters (Lohmann et al., 2004). The uncertain parameters are either considered as physical constants, or they are subject to calibration, which has been regarded as the established standard procedure to improve model behavior in numerous geoscientific fields.

Conventional calibration of distributed parameters in hydrologic models using gauged streamflow observation could pose an "ill-posed" challenge, primarily because of the large number of "free" parameters involved (Pokhrel et al., 2008; Refsgaard, 1997). Pixel-to-pixel calibration, on the other hand, typically produces artificial spatial discontinuities in parameter fields (Oubeidillah et al., 2014; Troy et al., 2008). Unreasonable parameterizations could lead to decreased physical realism and applicability of these process-based models. Estimating spatial fields of model parameters remains a big challenge in land surface and hydrologic modeling to advance the quest for physical realism (Archfield et al., 2015; Clark et al., 2016, 2017).

Obtaining spatially seamless parameter fields typically involves the implementation of spatial regularization, which offers significant advantages by reducing the dimensionality in the optimization while preserving the spatial parameter patterns. Spatial regularization typically comes in two forms: one involves the utilization of spatially constant parameter multipliers applied to predefined parameter fields, while the other employs spatially constant coefficients in transfer functions which link geophysical attributes to model parameters (Mizukami et al., 2017). The former approach necessitates a priori knowledge of parameter fields (Pokhrel & Gupta, 2010;

Methodology: Ruochen Sun, Baoxiang Pan
Supervision: Qingyun Duan
Validation: Ruochen Sun
Writing – original draft: Ruochen Sun
Writing – review & editing: Qingyun Duan

Sun et al., 2021), whereas the latter method directly calibrates global parameters using spatially distributed geophysical information (Beck et al., 2020; Hundecha & Bárdossy, 2004). Through the use of transfer functions, model parameters can be transferred across space, a process commonly known as parameter regionalization. For example, Samaniego et al. (2010) proposed the widely used multiscale parameter regionalization (MPR), which employs transfer functions at the native spatial resolution of geophysical data to rescale parameters to the desired spatial scale. However, the formulation of transfer function is usually uncertain. The primary limitation of MPR and other similar regionalization schemes lie in the selection of appropriate transfer functions (Samaniego et al., 2017), which could impact the effect of estimating parameters.

As deep learning (DL) techniques rapidly advance, they are garnering increasing interest within the geophysical community (Yu & Ma, 2021), presenting numerous opportunities and promising avenues to overcome obstacles (Reichstein et al., 2019), including improving model parameter generalization and performance (Shen et al., 2023). In terms of estimating spatially distributed parameter fields, Feigl et al. (2020) developed a method to automatically estimate transfer functions from geophysical data based on a text-generating neural network, which was implemented by using a variational autoencoder (VAE) architecture (Kingma & Welling, 2013). More specifically, the encoder of VAE consisted of word embeddings, convolutional layers and feedforward neural network (FNN) layers; the decoder of VAE involved a long short-term memory (LSTM) network and FNN layers. Feigl et al. (2022) further applied this method to the mesoscale Hydrologic Model using real-world data. Tsai et al. (2021) proposed a differentiable parameter learning (dPL) framework which used the neural network to map from raw input information to model parameters. This pioneering framework transforms the traditional model calibration problem into a parameter learning problem, harnessing the capabilities of modern DL computing infrastructure and the abundance of big data.

In land surface and hydrologic modeling, distributed models typically operate on regional or global scales, spanning multiple grid cells and several years, resulting in a substantial volume of output variables. System-scale performance metrics are typically composed of integrated spatial and temporal differences between simulated and observed variables. These performance metrics are widely employed as an objective function in conventional model calibration (Dembélé, Ceperley, et al., 2020). However, a key issue with system-scale performance metrics is their limited ability to fully exploit the rich information contained in the data (Clark et al., 2021). Conventional calibration involves condensing the wealth of information found in model outputs and observations into a single performance metric, and using the performance metric to estimate multiple parameters and various aspects of hydrological processes (Gupta et al., 2008). Such global calibration approaches can give rise to problems related to compensatory parameters, where the model may produce seemingly accurate results for the wrong reasons (Kirchner, 2006). When it comes to distributed modeling, conventional calibration methods can result in divergent, discontinuous, or even contradictory impacts on model predictions across different spatial domains. In simpler terms, one of the primary challenges frequently faced in the calibration of distributed models is that enhancements in one simulation region often come at the cost of reduced performance in others (Jiang et al., 2020; Sun et al., 2021; Xia et al., 2018; Yang et al., 2019). The aforementioned DL-assisted parameter estimation techniques excel in learning the relationship between inputs and model parameters. Nonetheless, they still heavily depend on system-scale performance metrics as loss functions for obtaining such relationship.

In this study, we introduce the Generative Adversarial Network-based Parameter Optimization (GAN-PO) method to learn spatially coherent parameter fields. The core aim of this method is harnessing the wealth of information within model outputs, observations, and other geophysical data to better enhance the precision and spatial consistency of model performance. We contend that adversarial learning (Goodfellow et al., 2014) represents a less-exploited yet highly appropriate framework for discerning the spatial characteristics of model simulation biases and subsequently estimating parameters. Through adopting an appropriate discriminative network for recognizing disparities between the simulations and observations, the insights gained from this discriminator are utilized to learn a transfer network which transforms the geophysical attributes to model parameters, thereby mitigating these discrepancies. We iteratively train both the discriminative network and the parameter transfer network to enhance the proficiency of each until the simulated output of the model becomes indistinguishable from the observed data. This adversarial learning framework presents the potential to fully exploit the information of observations and simulations to identify the model biases in a detailed manner, thus mitigating the dependence on system-scale performance metrics. We apply the GAN-PO method to learning parameters of the VIC (Hamman et al., 2018; Liang et al., 1994) land surface hydrologic model using ET as the

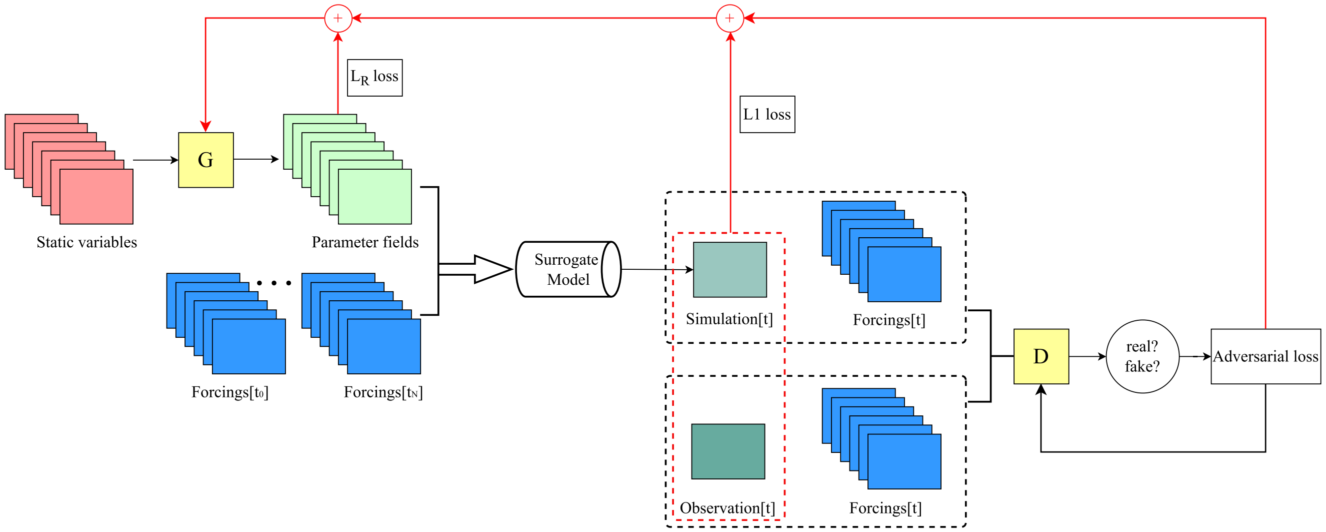

Figure 1. Schematic of the GAN-PO framework.

target variable over the Huaihe basin in China. For comparison, we incorporate two alternative methods: one being a traditional distributed calibration strategy employing the MPR technique, and the other being the state-of-the-art dPL method.

The rest part of the paper is organized as follows. Section 2 describes the detailed methodology. Section 3 briefly describes the model, the study region, data used in this study, the experiment setup, and the evaluation metrics. Section 4 presents the detailed results, followed by discussion in Section 5. Section 6 provides the conclusions of this paper.

2. Methods

2.1. Generative Adversarial Network-Based Parameter Optimization (GAN-PO)

2.1.1. General Framework

The precise depiction of spatial patterns is a distinctive and vital characteristic of distributed land surface hydrologic models. Nevertheless, the spatial predictability of these models faces significant challenges due to the prevalent evaluation framework that centers on aggregated performance metrics. The challenge is to untangle errors associated with various domains from these metrics.

To address this challenge, we introduce the GAN-PO method. GAN, which was initially developed by Goodfellow et al. (2014), is a class of DL framework for generative artificial intelligence. In GANs, two neural networks engage in a zero-sum game, where a gain for one agent corresponds to a loss for the other. Specifically speaking, a neural network, referred to as the generator (G), is designed to learn a mapping from a latent space to a data distribution of interest. In parallel, the other neural network, termed the discriminator (D), plays a role in guiding the training of G by discerning candidates generated by G from the target distribution. The concept of adversarial learning signifies that G is not trained toward a fixed, predefined objective but rather to deceive D, which in turn is iteratively updated to enhance its discrimination capability.

In the proposed GAN-PO method (Figure 1), we apply the generator G as a deterministic transfer function to convert distributed fields of geophysical attributes x into model parameter fields $G(x)$. Subsequently, these parameter fields serve as inputs to the model $h()$ for simulating variables of interest. It is essential to note that the model $h()$ must be differentiable, enabling the tracking of gradients (Shen et al., 2023). A differentiable surrogate model can be used to replace the original physical model for convenience. The discriminator D is trained to distinguish whether a sample is from observations Y or simulations Y'. Y' is directly related to $G(x)$, which can be succinctly expressed as $Y' = h(G(x))$, in which other model inputs are not listed. The training objective of the generator is to fool the discriminator by generating novel parameters in a way that the discriminator perceives the resulting simulations as observed. These can be achieved by applying the adversarial loss:

$$L_{GAN}(G,D) = E_y[\log D(y)] + E_x[\log(1 - D(h[G(x)])] \tag{1}$$

where G tries to minimize this loss against an adversarial D that tries to maximize it, that is, $min_G max_D L_{GAN}$ (G, D). When $h()$ is a DL-based surrogate model, the weights are fixed and excluded from updating through back-propagation, while the gradient information is allowed to pass through.

The adversarial loss listed above is a general objective for GAN training, which can be fulfilled by several solutions. However, adversarial loss alone may lead to instability during training because the generator and discriminator are in a continuous competition to outperform each other. This can cause the training to collapse, where the generator fails to accurately represent the data. In addition, previous studies (Isola et al., 2017; Ravuri et al., 2021) have also suggested the advantages of combining the adversarial loss with a more traditional loss, such as L1 distance. In this case, the generator's objective extends beyond merely deceiving the discriminator; it also aims to approximate the ground truth output, which is important for the model to produce accurate predictions. Therefore, when the GAN model is used in parameter calibration, the inclusion of certain regularization terms becomes necessary. In this study, we choose to use L1 norm as a regularization term, following recommendations from prior studies (Isola et al., 2017; Pan et al., 2021; Ravuri et al., 2021). L1 norm is less sensitive to outliers than L2 norm because it penalizes the absolute value of the weights rather than their squares. This can make the model more robust to noise and unusual data points, which is beneficial for the stability of GAN training. We also introduce a regularization term L_R that penalizes generated parameter values outside their respective ranges. L_R is expressed as follows:

$$\begin{aligned} L_R &= L_{ub} + L_{lb} \\ L_{ub} &= \begin{cases} \dfrac{sum[\mathrm{ReLU}(G(x) - ub)]}{sum[\mathrm{ReLU}(G(x) - ub) > 0]}, & sum[\mathrm{ReLU}(G(x) - ub) > 0] \neq 0 \\ 0, & sum[\mathrm{ReLU}(G(x) - ub) > 0] = 0 \end{cases} \\ L_{lb} &= \begin{cases} \dfrac{sum[\mathrm{ReLU}(lb - G(x))]}{sum[\mathrm{ReLU}(lb - G(x)) > 0]}, & sum[\mathrm{ReLU}(lb - G(x)) > 0] \neq 0 \\ 0, & sum[\mathrm{ReLU}(lb - G(x)) > 0] = 0 \end{cases} \end{aligned} \tag{2}$$

where lb and ub represent the lower and upper bounds of the parameter ranges. L_{ub} represents the average value of generated parameter values exceeding the upper bound of the parameter ranges, while L_{lb} signifies the opposite scenario. Minimizing L_R pushes generated parameter values to fall within their specified parameter ranges.

Furthermore, drawing inspiration from Pan et al. (2021), the discriminator takes into account both the forcings and output, thereby enabling discrimination between different weather conditions. The final objective function of the GAN-PO method is:

$$\begin{aligned} L_{GAN-PO} = {} & E_y[\log D(f,y)] + E_x[\log(1 - D(f,h[G(x)])] \\ & + \lambda_l E_{x,y}[\|h[G(x)] - y\|_1] \\ & + \lambda_r L_R \end{aligned} \tag{3}$$

where f denotes forcings of the model, λ_l and λ_r are weights controlling the relative importance of the regularization terms; $\mathrm{ReLU}(x) = max(0,x)$.

2.1.2. Model Configuration

The architectural framework of the GAN-PO method is elucidated in Figure 2. For the generator of GAN-PO, a U-net form (Ronneberger et al., 2015) of convolutional neural network (CNN) is applied. The U-net integrates a convolution-based contracting path to capture input field information. Simultaneously, it features a symmetric transposed convolution-based expanding path, gradually refining output. Inclusion of skip connections between symmetrical convolution and transposed convolution blocks facilitates the direct transfer of information across the network. By doing so, skip connections contribute to the model's ability to retain and refine detailed

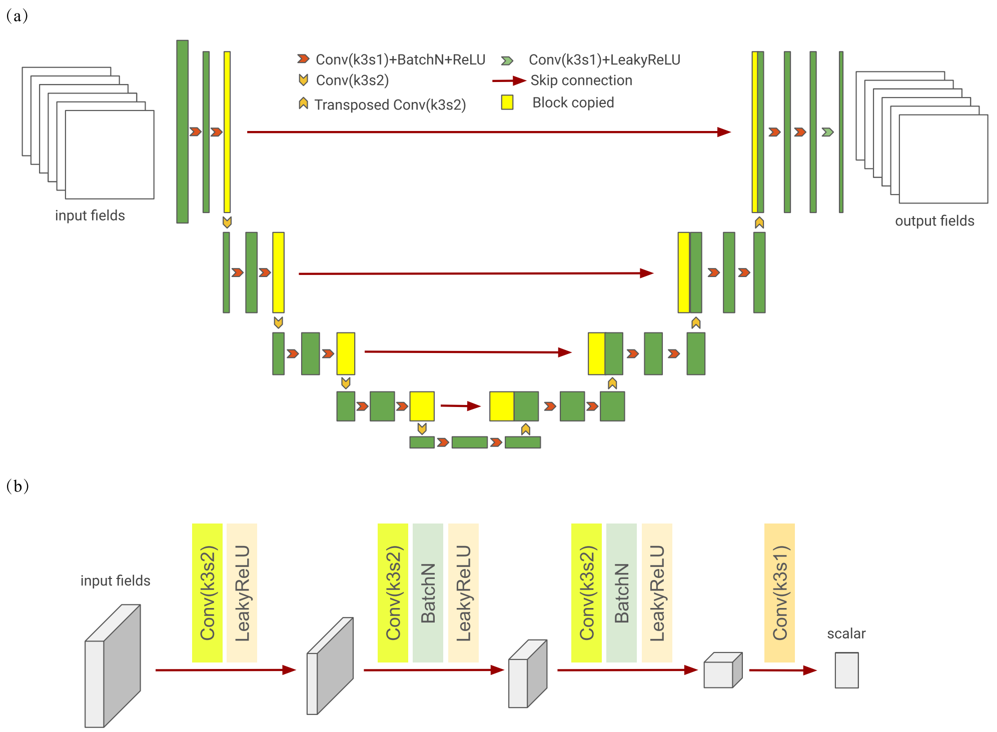

Figure 2. Architecture of (a) generator and (b) discriminator network in the GAN-PO method. k and s denote the kernel size and stride of the convolutional layer.

information, ultimately improving the overall performance and training stability of the network. For the discriminator of GAN-PO, we utilize the foundational architecture of deep convolutional GANs (DCGANs), following the guidelines delineated by Radford et al. (2015). More specifically, the discriminator receives input fields, which are then processed through a convolutional layer and a LeakyReLU activation function, followed by a sequence of convolution-BatchNorm-LeakyReLU layers with 3 × 3 filters and a stride of 2. Subsequently, a convolutional layer is applied to generate a 1-dimensional output after the final layer. We use LeakyReLU activation with a parameter of $\alpha = 0.2$, to avoid max-pooling throughout the network. The detailed architectures of the generator and discriminator in GAN-PO can be found in Table S1 and Table S2 in Supporting Information S1, respectively.

To facilitate the training of the GAN-PO model, we implement specific techniques aimed at stabilizing the training procedure. For L_{GAN} (as described in Equation 2), we opt for increased stability during training by replacing the negative log likelihood objective with a least squares loss, a technique introduced by Mao et al. (2017). This modification enhances training stability and produces higher-quality results. Specifically, for the L_{GAN} loss, we train the generator to minimize the expression $E_x[(D(f,h[G(x)]) - 1)^2]$ and train the discriminator to minimize $E_y[(D(f,y) - 1)^2] + E_x[D(f,h[G(x)])^2]$.

In contrast to supervised learning, where training concludes upon reaching a minimum loss for the validation set, determining the optimal endpoint for adversarial training lacks consensus. A common practice in computer vision when training GANs involves visually inspecting generated results and halting training if no discernible improvement is observed. However, this strategy proves inadequate for our problem, as visual judgments of ET fields can be easily deceived, leading to suboptimal quantitative performance during testing. To address this challenge, we systematically calculate crucial indices, which are mean square error (MSE) and Kling-Gupta efficiency (KGE; Gupta et al., 2009) in this study, between simulations and observations for the validation set at the end of each training epoch. The generator is saved at instances where these statistical measures demonstrate the most favorable alignment.

2.2. Baseline Methods

We adopt two kinds of distributed parameter calibration methods as baselines. One is the widely used MPR technique, in which the spatially constant coefficients of transfer functions are estimated by the shuffled complex evolution (SCE-UA) algorithm (Duan et al., 1992). The other is the recently developed dPL (Tsai et al., 2021) framework.

As illustrated in the workflow diagram depicted in Figure S1 in Supporting Information S1, the conventional calibration method in this study adopts the MPR technique, which encompasses parameter regionalization based on transfer functions, parameter scaling. Subsequently, model parameters are estimated through the adjustment of global transfer function parameters, rather than directly modifying the model parameters themselves. We use the SCE-UA algorithm to optimize the global transfer function parameters. Reference for the transfer functions and upscaling operators for the VIC model parameters that require calibration can be found in Table 1 of Gou et al. (2021).

dPL is a deep neural networks-based framework which can learn a global mapping between inputs and model parameters. There exist two versions of dPL (g_A and g_z), each tailored for specific use cases within the field of geosciences. The dPL framework incorporates a parameter estimation module that facilitates the mapping from raw input information to model parameters. This input information can either be observable attributes for g_A or forcing-response pairs for g_z. LSTM was chosen as the network structure for the mapping in the original study. Therefore, information derived from observations of a grid cell can be readily propagated to model parameters of that cell through the well-trained network. These parameters are subsequently input into a differentiable model or, alternatively, a surrogate model such as a neural network. The loss function is defined over the entire training data set by combining the sum of squared differences for all sites at once. dPL was implemented on the distributed VIC model to estimate parameters using soil moisture and streamflow as targets in two cases. Interested readers can refer to Tsai et al. (2021) for more information.

3. Experimental Design

3.1. Model and Data

The study area is the same as that in our previous study (Sun et al., 2023), which is a rectangle region covering the upper Huaihe basin (Figure 3). The widely used VIC model is used in this study. The VIC model is operated at a spatial resolution of 0.25°, resulting in the study region of 24 × 17 grid cells. The data used in this study consist of the forcing data of the VIC model, the observed target variable for calibration, and the static geophysical attribute data. The VIC forcings come from the China Meteorological Forcing Data set (CMFD) data set with a temporal resolution of 3 hours and a spatial resolution of 0.1° (He et al., 2020). All meteorological variables are processed to achieve a daily temporal resolution and a spatial resolution of 0.25°. The target variable used for calibration in this study is evapotranspiration (ET). We use the Global Land Evaporation Amsterdam Model (GLEAM) V3.7 ET product (Martens et al., 2017) as reference, which has been shown to exhibit relatively high accuracy (Jia et al., 2022; Mao et al., 2024). GLEAM ET data set is available on a 0.25° regular grid and at daily temporal resolution. The static geophysical features include porosity, bulk density, soil profile depth as well as sand, silt, and clay percentages from the China Data set of Soil Hydraulic Parameters (Dai et al., 2013) and the Soil Database of China for Land Surface Modeling (Shangguan et al., 2013).

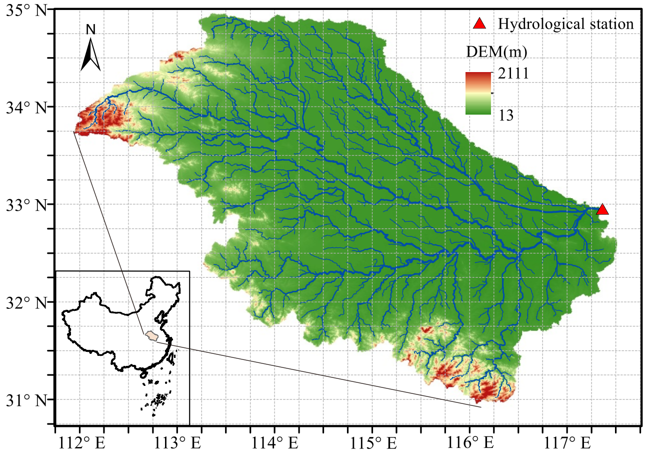

Figure 3. The rectangular study area encompassing the upper Huaihe basin. Each small box, delineated by intersecting dashed lines, represents a 0.25° × 0.25° grid cell.

3.2. Application of Different Methods to Parameter Estimations

The VIC parameters needed to be calibrated include the variable infiltration curve parameter (b), maximum velocity of baseflow (Dsmax), fraction of Dsmax where non-linear baseflow begins (Ds), fraction of maximum soil moisture where non-linear baseflow occurs (Ws), the second soil layer thickness (d2), the third soil layer thickness (d3), parameter characterizing the variation of saturated hydraulic conductivity (E). The time periods 2009–

2010, 2013–2014, and 2011 are designated for training, validation, and testing for all three methods.

For the conventional calibration method, the transfer functions and upscaling methods used in MPR for the VIC model are referred to Gou et al. (2021). In the context of a fully distributed model calibration problem, where both the observed and modeled variables are represented by 3D arrays, the calculation of the objective function directly influences the calibration results. Dembélé, Ceperley, et al. (2020) evaluated the model performances across four distinct calibration strategies when ET was used as a calibration variable. Based on the conclusions drawn from Dembélé, Ceperley, et al. (2020), we adopt the spatial bias-accounting strategy, which achieved the highest improvement in the overall model performance, for calibrating the VIC model. In the spatial bias-accounting strategy, the objective function is defined by matching the modeled and observed ET for all grid cells at each time step. More specifically, root mean square error (RMSE) is computed at each time step by comparing the observed and modeled ET across all grids of the region. The objective function to be minimized is the average of the RMSE calculated for all time steps. When employing SCE-UA for optimization, the convergence criteria are established as follows: either the objective function value remains within a 1% change range over 10 consecutive iterations, or the number of model evaluations reaches the maximum allowable number of 2000.

For both the GAN-PO and dPL methods, we employ the Adam optimizer (Kingma & Ba, 2014) with identical learning rate of 1e−4 and identical momentum parameters ($\beta_1 = 0.5$, $\beta_2 = 0.999$), and we set the maximum training epochs to 200. In addition, we employ batch sizes of 32 and 50 for GAN-PO and dPL, respectively. In the dPL method, we randomly select grid cells and time periods with a length of 30 days within the training data set to compose a minibatch for training, aligning with the work of Tsai et al. (2021). Both models are trained on a NVIDIA A100 GPU with 40 GB video memory. The crucial hyperparameters in the GAN-PO method are λ_l and λ_r. Since LSTM was used in dPL as the parameter estimation network, which maps raw data to model parameters, the hyperparameter in the dPL method is the hidden size of the LSTM. In our study, hyperparameters are determined using a grid search method over a range of parameter values. The final configuration is selected by taking the parameter set that achieves the highest performance metric during the validation period among all possible combinations. We consider the potential values of λ_l to be: 0.5, 1, 10, 20, 30, 40; the potential values of λ_r to be: 0.1, 0.5, 1, 1.5, 2. In addition, we consider the possible values of hidden size of LSTM to be: 64, 96, 128, 196, 256. Finally, λ_l and λ_r are set to 30 and 1, respectively, and the hidden size is configured to be 128.

In the original dPL paper, LSTM was used as a surrogate model to approximate VIC while allowing for gradient tracking. The grid-level training mode of dPL also has the advantage of adjusting the sampling density to balance model behavior and training time. To ensure fairness in the comparison, we first train a LSTM surrogate, utilizing seven meteorological forcing variables and seven model parameters as input, with simulated ET as the target. The training procedure resembles that of our previous study (Sun et al., 2023). Specifically, we randomly generate 400 distributed fields of the VIC parameters and run the VIC model 400 times to obtain simulated ET. The LSTM model is trained using 200 input-output data pairs, with another 100 data pairs for validation. Given that the input-output data of each grid cell can be utilized for training, we have a total of 81,600 (200 × 408) grid cells, each containing 2 years of data, employed for training. The LSTM surrogate has a hidden size of 128 and a single fully connected layer. We train the model using a batch size of 400 and MSE loss, incorporating the early stopping technique. The LSTM operates in a sequence-to-value mode with a sequence length of 30 days. Consequently, predicting a daily ET value necessitates meteorological forcings from the preceding 29 days, in addition to the forcing data of the target day. Similar to Tsai et al. (2021), we dynamically update the LSTM surrogate model during calibration by incorporating additional sets of parameter samples. We evaluated the performance of the LSTM surrogate in replicating VIC simulated results. Figure S2 in Supporting Information S1 shows that the overall accuracy of the final trained surrogate model is satisfactory. This LSTM surrogate is then employed consistently across all three methods.

Our previous study (Sun et al., 2023) demonstrated that the LSTM surrogate may not reliably ensure the accurate representation of spatial patterns in distributed models. Instead, we employed a deep autoregressive neural network (ARnet), which converts the distributed surrogate modeling to an image-to-image regression task, to construct a dependable surrogate system for the VIC model. Given that the ARnet surrogate directly takes parameter fields as input and outputs target field, it could be more suitable for the GAN-PO method. We conduct further investigation to ascertain whether the incorporation of the ARnet surrogate can improve optimization performance. The detailed architecture of the ARnet can be referenced in Sun et al. (2023). The training process closely resembles that of our previous study, with the exception that the training target is replaced with ET. Unless otherwise specified, the LSTM surrogate is used in the GAN-PO method.

3.3. Evaluation Metrics

We utilize various metrics to assess the performance of the model, including point-to-point and spatial evaluations, for both the baselines and the GAN-PO method on the test data set. For grid scale evaluation, we calculate RMSE, KGE and correlation coefficient (CC) between observation and simulation in each grid cell. For spatial evaluation, we use bias-insensitive spatial performance metrics, namely empirical orthogonal function (EOF) analysis, fractions skill score (FSS) and structural similarity index (SSIM), to measure the similarity between the simulation and observation patterns.

3.3.1. Common Statistic Metrics

Three widely employed statistical metrics for evaluating model performance at the grid scale are the RMSE, CC and KGE. The formulas are given by:

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{N}\left(S_j^i - O_j^i\right)^2}{N}} \tag{4}$$

$$CC_j = \frac{\sum_{i=1}^{N}\left(O_j^i - \overline{O_j}\right)\left(S_j^i - \overline{S_j}\right)}{\sqrt{\sum_{i=1}^{N}\left(O_j^i - \overline{O_j}\right)^2}\sqrt{\sum_{i=1}^{n}\left(S_j^i - \overline{S_j}\right)^2}} \tag{5}$$

$$KGE_j = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{6}$$

$$\beta = \frac{\overline{S_j}}{\overline{O_j}}$$

$$\gamma = \frac{\sigma_j^s}{\sigma_j^o}$$

where S_j^i and O_j^i are the simulated and observed ET in the i th day on the j th grid, $\overline{S_j}$ and $\overline{O_j}$ are their corresponding mean values, respectively. N is the total number of events. In the KGE formulation, r denotes the linear correlation coefficient, β and γ represent the bias and relative variability in the simulated and observed values. σ_j^s and σ_j^o are the standard deviation of simulated and observed variables on the j th grid, respectively.

To facilitate a more in-depth comparison of model errors between the utilization of calibrated parameters and default parameters, we introduce the concept of "relative RMSE change." This metric is defined as follows:

$$\Delta = \frac{RMSE_d - RMSE_c}{RMSE_d} \times 100\% \tag{7}$$

where $RMSE_c$ and $RMSE_d$ are the RMSEs when the model is run with calibrated and default parameters, respectively. This metric is valuable for assessing the impact of parameter calibration on model performance and understanding the relative improvement or degradation in predictive accuracy. Positive value of Δ means that model error is reduced after calibration.

3.3.2. Spatial Evaluation Methods

In this study, we incorporate bias-insensitive spatial performance metrics, namely EOF, FSS and SSIM. These metrics are employed to quantify the resemblance between simulation and observational patterns, providing a comprehensive assessment of spatial model performance. The use of bias-insensitive metrics is advantageous in capturing not only the accuracy but also the spatial distribution and patterns of simulated and observed data.

The EOF method is a popular technique used in fields such as atmosphere, climate, ocean, and hydrology science to identify potential spatial patterns of variability and how they change over time. Koch et al. (2015) introduced a novel approach of performing a joint EOF analysis on a combined data matrix containing both reference and simulated data. This approach allows for the identification of spatiotemporal patterns of variability for both data sets, and enables the calculation of a similarity score between them. The difference between the loadings at each time step can serve as an indicator of spatial similarity, with the loading deviation weighted according to the corresponding EOF's variance contribution to ensure a reliable pattern similarity score. The EOF-based similarity score between a reference map and a modeled map at time t can be formulated as follows:

$$S^{t}_{\mathrm{EOF}} = \sum_{i=1}^{n} w_i \left| \left(load_i^s - load_i^o \right) \right| \tag{8}$$

where w_i is the variance contribution of the i th EOF, n is the total number of orthogonal modes (EOFs), $load_i^s$ and $load_i^o$ are the corresponding loadings of simulated and observed maps, respectively. Smaller score indicates a higher degree of similarity between the reference map and the modeled map.

The FSS is a scale-selective performance metric developed by (Roberts & Lean, 2008) for assessing the accuracy of precipitation forecasts at different spatial scales, for a given threshold. It uses the nearest neighbor approach to select the relevant scales for analysis. FSS calculates the fractional coverage of binary events that exceed the defined threshold, within a given spatial window. Typically, percentile thresholds are used to convert continuous data into binary fields to remove the impact of bias and focus on spatial accuracy. The main steps for obtaining the FSS are as follows: (a) convert the reference and modeled spatial patterns into binary fields for a specific threshold; (b) for each grid in each binary field, compute the fraction of grids with a value of 1 within a given square window of length n; (c) calculate the MSE between the referenced and modeled fraction fields; (d) obtain the final FSS by normalizing the MSE from step 3 with the largest possible MSE that can be obtained from the modeled and referenced fractions. The FSS at a spatial scale of n, for a given threshold, can be expressed as:

$$FSS_{(n)} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} \left(O_{(n)i} - S_{(n)i} \right)^2}{\frac{1}{N}\left(\sum_{i=1}^{N} O_{(n)i}^2 + \sum_{i=1}^{N} S_{(n)i}^2 \right)} \tag{9}$$

where $O_{(n)}$ and $S_{(n)}$ are the resultant fields of referenced fractions and modeled fractions, respectively. N is the total number of valid grids in the domain. The FSS ranges from 0 (complete mismatch) to one (perfect match).

SSIM is used for measuring the similarity between two images. SSIM operates as a perception-based model that evaluates image degradation by considering perceived changes in structural information. It incorporates essential perceptual phenomena, including luminance masking and contrast masking terms. The SSIM for two images x and y is defined as:

$$\mathrm{SSIM}(x,y) = \frac{2\left(\mu_x \mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{10}$$

where μ_x and μ_y denote means of two images, σ_x^2, σ_y^2 represent variances of the two images, and σ_{xy} are their covariance. In practical applications, means and standard deviations are commonly computed using a sliding Gaussian window. The SSIM index is a numeric value in the range of -1 to 1. A score of 1 signifies perfect similarity between the compared images, 0 indicates no similarity, and -1 indicates perfect anti-correlation.

4. Results

We first conduct a comparison study of the three calibration methods, employing the same LSTM surrogate model to ensure fairness. Our emphasis is on commonly utilized statistical metrics at the grid scale, along with

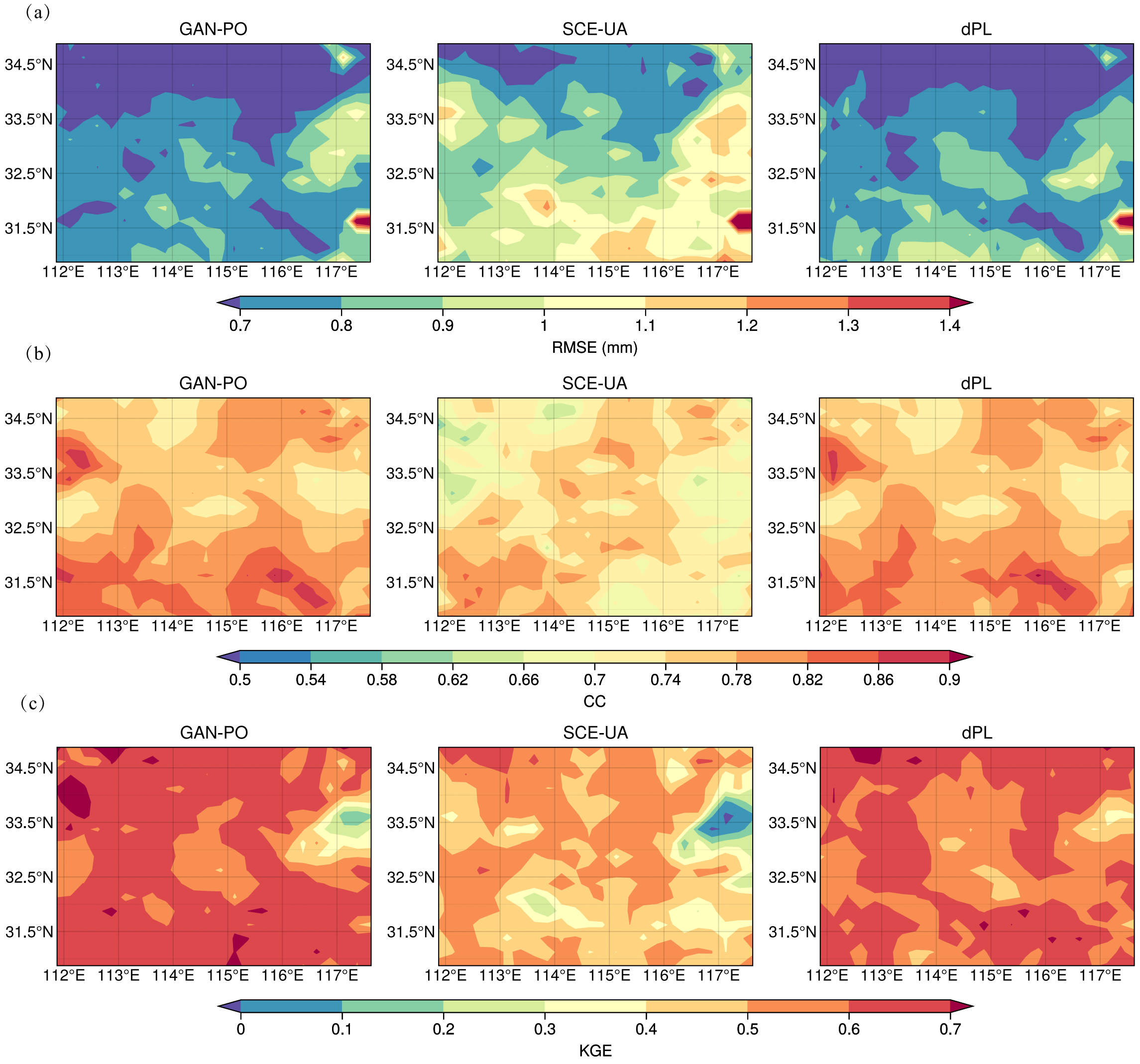

Figure 4. Spatial distributions of (a) RMSE, (b) CC and (c) KGE values between observations and simulated ET by the surrogate model using parameters calibrated by different methods.

the assessment of spatial coherence at the field scale. We additionally conduct a thorough analysis to investigate whether the incorporation of a more suitable surrogate in GAN-PO can enhance optimization performance.

4.1. Comparison of Optimization Performance at Grid Scale

Figure 4 shows the spatial maps of RMSE, CC and KGE values for simulated ET using parameters calibrated by different methods. Generally, the model performance can be improved by all three methods. GAN-PO and dPL demonstrate comparable performance at the grid scale, and they outperform SCE-UA, demonstrating notably reduced RMSE values as well as increased CC and KGE values. Approximately 80% and 77% of grid cells exhibit a CC greater than 0.75 for GAN-PO and dPL simulated ET, respectively. In contrast, simulated ET from SCE-UA has CC values greater than 0.75 in only 37% of grid cells. Furthermore, the percentage of grid cells with KGE values larger than 0.6 for GAN-PO, dPL and SCE-UA is 73%, 56% and 4.2%, respectively. Both SCE-UA and dPL employ RMSE as objective functions, defined over the entire training data set. The distinction lies in the fact that dPL calculates RMSE on the minibatch, while SCE-UA utilizes all available years of training data to calculate RMSE. It's worth noting that despite not directly employing RMSE during training, GAN-PO still demonstrates substantial improvements across a diverse range of regions.

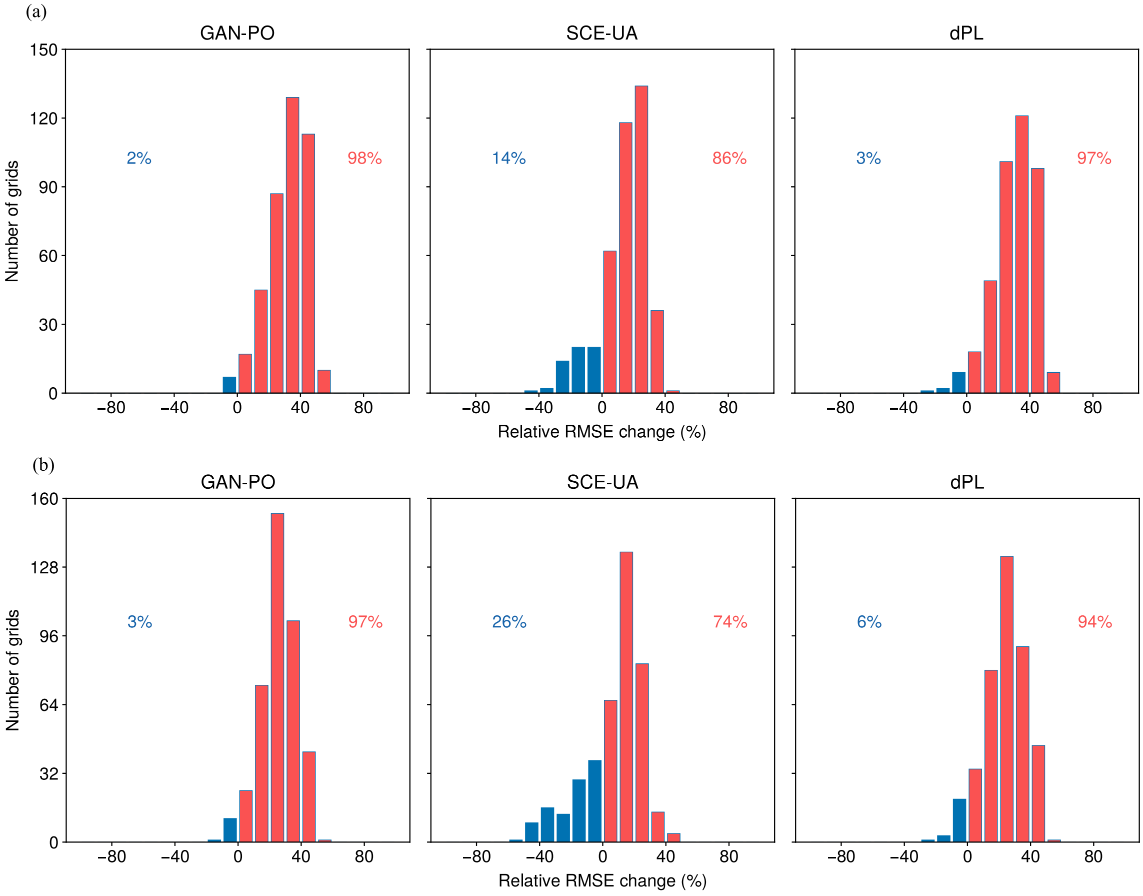

Figure 5. The histograms of the relative change (Δ) in RMSE for simulated ET by the surrogate model over the (a) training period and (b) testing period under various calibration scenarios. Red bars indicate a reduction in the RMSE of simulated ET when calibrated compared to using default parameters. Conversely, blue bars signify an increase in RMSE after calibration.

We further quantitatively investigate the changes in model performance using the optimized parameters compared to that using the default parameters. Figure 5 depicts the distributions of the relative RMSE change Δ (Equation 5) for ET. During the testing period, GAN-PO reduces the RMSE of simulated ET for nearly the entire region (97% of the total grid cells), with minimal degradation observed from the training period. dPL can reduce the RMSE in 94% of the total grid cells, showing a slightly inferior performance compared to GAN-PO but maintaining robust performance for both training and testing sets. In contrast, SCE-UA results in a certain degree of degradation in model performance, with a notable worsening observed during the testing period, indicating the temporal generalization of calibrated parameters are poor. Similar results were observed in our previous study (Sun et al., 2021), suggesting that conventional calibration against a single performance metric aggregated over time and domain may cause the problem of compensatory parameters. Specifically, parameters within some specific grids may be assigned unreasonable values which compensate for unrealistic parameter values in other grids.

4.2. Comparison of Optimization Performance at Field Scale

This study incorporates spatial evaluations utilizing bias-insensitive spatial performance methods. Spatial evaluations are essential in the realm of distributed land surface hydrologic modeling because a primary objective is to attain uniform and spatially consistent model performance across diverse regions. In Figure 6a, the EOF-based similarity scores for different calibration methods are presented. Lower values of the EOF-based metric are preferable, indicating a high degree of similarity in the spatial pattern between the model-predicted ET and observed ET. The time series of EOF-based scores exhibit some fluctuation, notably less pronounced in winter and more obvious in summer, potentially attributed to the limited spatial variability of ET in winter. It is evident that GAN-PO demonstrates superior spatial performance compared to dPL and SCE-UA. Additionally, we also adopt the popular SSIM to assess the level of similarity between two ET fields. Consistent results, which are shown in Figure 6b, can be achieved in relation to the SSIM series during the testing period. While GAN-PO generally exhibits superior spatial performance, as evidenced by the lowest mean EOF-based similarity scores and the highest mean SSIM values, it is worth noting that dPL outperforms GAN-PO in specific periods. Upon closer examination, we observe that dPL primarily outperforms GAN-PO during periods characterized by high ET values, particularly between June to September. One possible reason for this observation is that dPL utilizes RMSE as the loss function, which places more emphasis on accurately estimating high ET values. Additionally, dPL constructs minibatches for training by explicitly utilizing ET series from various grid cells with a length of 30 days, whereas GAN-PO incorporates time information implicitly through the inclusion of forcings into the discriminator input. Consequently, during periods characterized by significant variations in ET patterns, dPL may capture rapid changes in ET more effectively, whereas GAN-PO may be less responsive to such rapid changes.

Furthermore, FSS, a scale-dependent verification method, is utilized in our study. Figure 7 illustrates FSS curves for average ET patterns during the testing period across various percentile thresholds. Higher FSS values are preferred, and the curves consistently increase with the scale used for fraction computation, aligning with the method's inherent characteristics. Notably, SCE-UA and dPL exhibit similar proficiency in capturing low ET patterns. However, SCE-UA displays diminished performance compared to dPL for high thresholds. In contrast, GAN-PO demonstrates superior spatial performance overall. It has the best skill in reproducing localized features within regions characterized by low ET values and surpasses the performance of the other two methods in modeling the highest 5% ET patterns. For other high

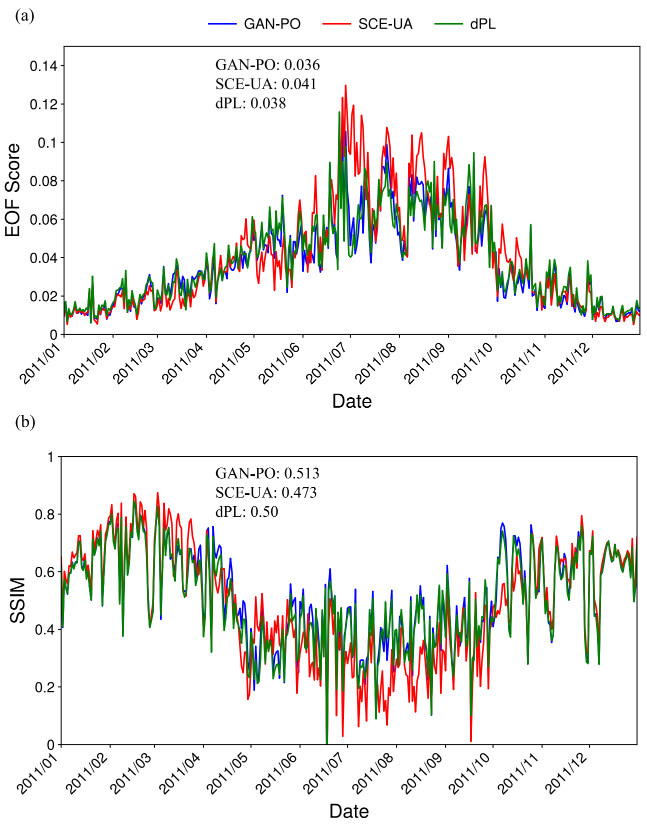

Figure 6. Time series comparison of (a) EOF-based similarity scores and (b) SSIM values for different calibration methods during the testing period. The average values over time are also displayed.

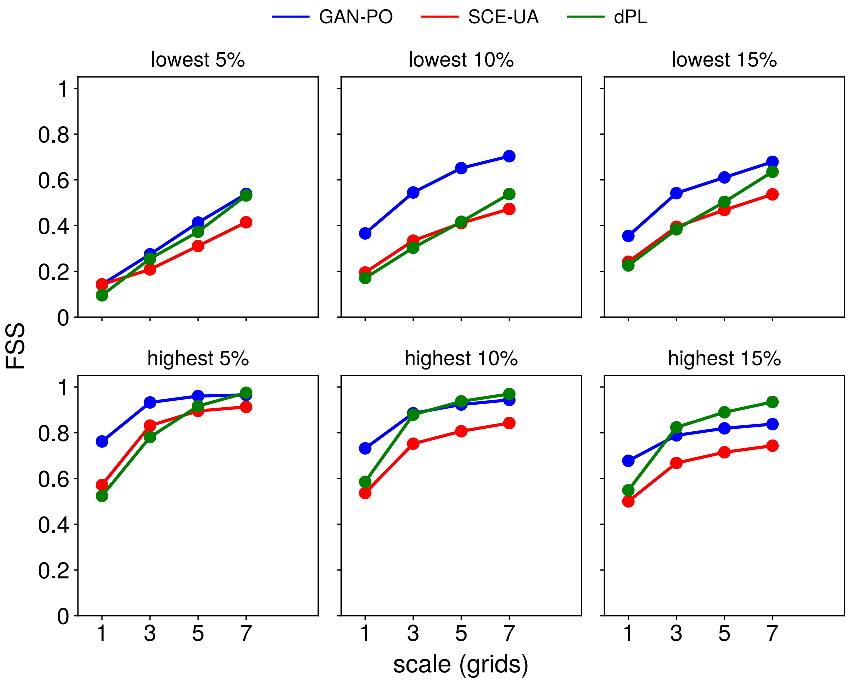

Figure 7. Plots of FSS against neighborhood length for the averaged ET patterns over the testing period utilizing different calibration methods.

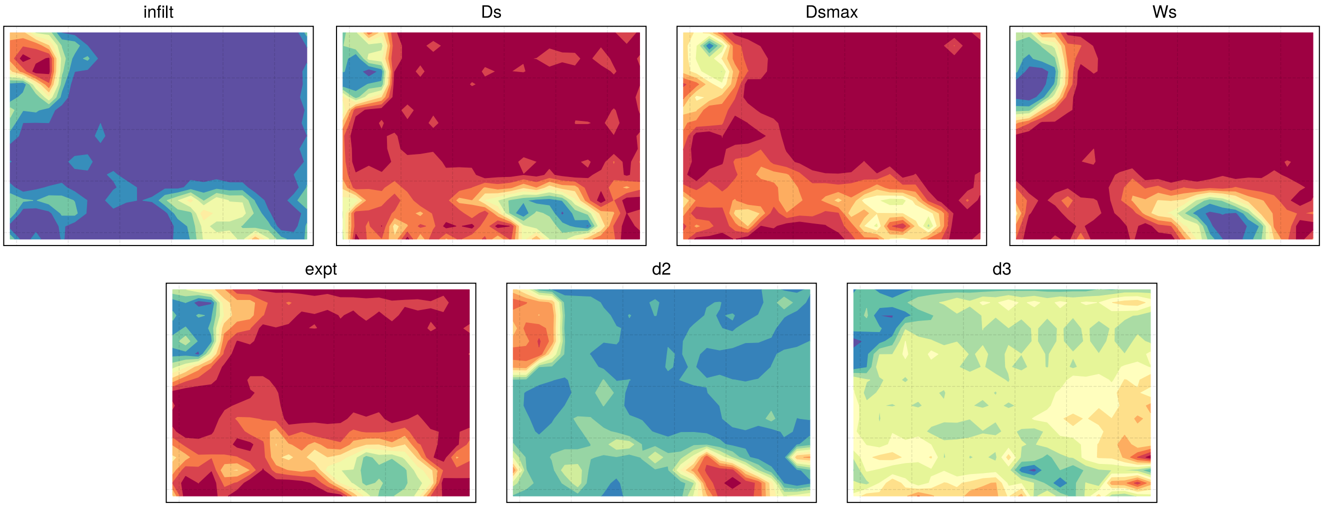

Figure 8. Spatial distribution of GAN-PO estimated parameters.

value thresholds, dPL exhibits a slight performance advantage over GAN-PO, which aligns with the results and analysis presented above.

The superior spatial performance of GAN-PO implies that the generator has successfully acquired a reasonable parameter mapping from static geophysical attributes. The spatial pattern of the estimated parameters exhibits a tendency to align with the topographic pattern of the study region (Figure 8). The main reason behind the advantage of GAN-PO lies in its ability to enable the discriminator to thoroughly explore spatial features and hidden information within both the observation and simulation domains. This unique capability enhances the method's capacity to discern intricate patterns and relationships within both observed and simulated data, contributing to more effective and informed parameter generative processes.

4.3. Ablation Analysis

To demonstrate the indispensability of the discriminator term in the GAN-PO method, we perform an ablation study to isolate the effect of the L1 loss and the adversarial loss L_{GAN} in Equation 2. If only the L1 loss is retained, the discriminator is eliminated, and the GAN-PO method turns into a similar form to that of dPL, albeit with a distinction: the generator within GAN-PO directly integrates distributed fields of geophysical attributes as its input. We refer to "L1" as the GAN-PO method without the discriminator term. Figure 9 and Table 1 shows the evaluation results of the "L1". Utilizing the L1 loss alone in the GAN-PO method for parameter calibration yields significantly deteriorated performance for both grid-scale evaluation and spatial evaluation. Although it still performs slightly better than the SCE-UA method, relying solely on the traditional loss cannot enable the method to discern intricate patterns and relationships within both observed and simulated data. This limitation not only leads to poor model capacity for predicting spatial patterns but also results in inconsistent model behavior over a large domain. Similar observations have been reported in previous literature on GAN-related image translation (Isola et al., 2017).

4.4. Importance of the Appropriate Surrogate Models

As mentioned in Section 3.2, the ARnet surrogate may be better suited for the GAN-PO method, as it adopts an image-to-image regression strategy. Therefore, we can directly feed the meteorological forcing fields and the parameter fields generated by the generator of GAN-PO into ARnet without the need for grid discretization. We use the terms GAN-PO_LSTM and GAN-PO_ARnet to distinguish between the utilized surrogate models. It is clear from Figure 10 that the GAN-PO method can further reduce the RMSE and increase the CC when the ARnet is utilized as the surrogate model. GAN-PO_ARnet generates simulated ET with superior spatial performance,

Table 1
Evaluation Metrics for the GAN-PO Method When Employing Different Loss Functions

Loss	RMSE	EOF score	SSIM
L1	0.849	0.0387	0.475
L1 + L_{GAN}	0.737	0.036	0.513

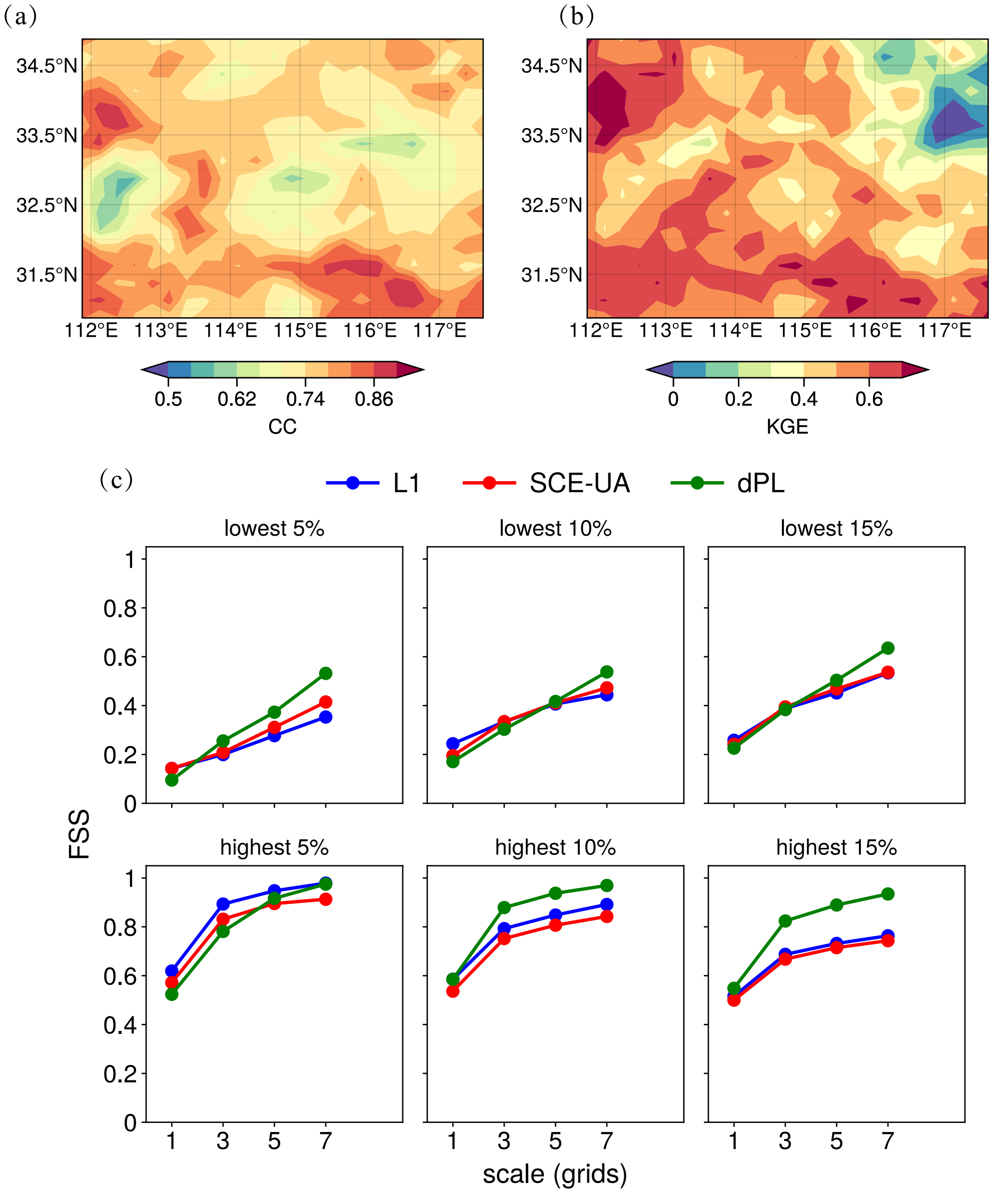

Figure 9. Evaluation results of the simulated ET by the surrogate model using parameters calibrated by the GAN-PO method without inclusion of the discriminator. (a) Spatial map of CC, (b) spatial map of KGE and (c) FSS compared to other calibration methods.

characterized by higher FSS values, except under the highest 5% threshold. In addition, the average EOF similarity score and SSIM value of GAN-PO_ARnet show a 10.8% and 5.8% improvement, respectively, in comparison to GAN-PO_LSTM. Considering that both the generator and discriminator components in GAN-PO exploit the strengths of convolutional neural networks in high-dimensional image processing and spatial feature extraction, employing ARnet as a surrogate has the potential to improve the retention of spatial information within the parameter fields learned from geophysical attributes, consequently enhancing the spatial performance of the model.

5. Discussion

In this study, we propose the GAN-PO method for distributed model calibration. We incorporate two regularization terms into the loss function for training the generator. Figure S3 in Supporting Information S1 displays the time series tracking various errors in the loss function during the training of the generator at each epoch. It can be

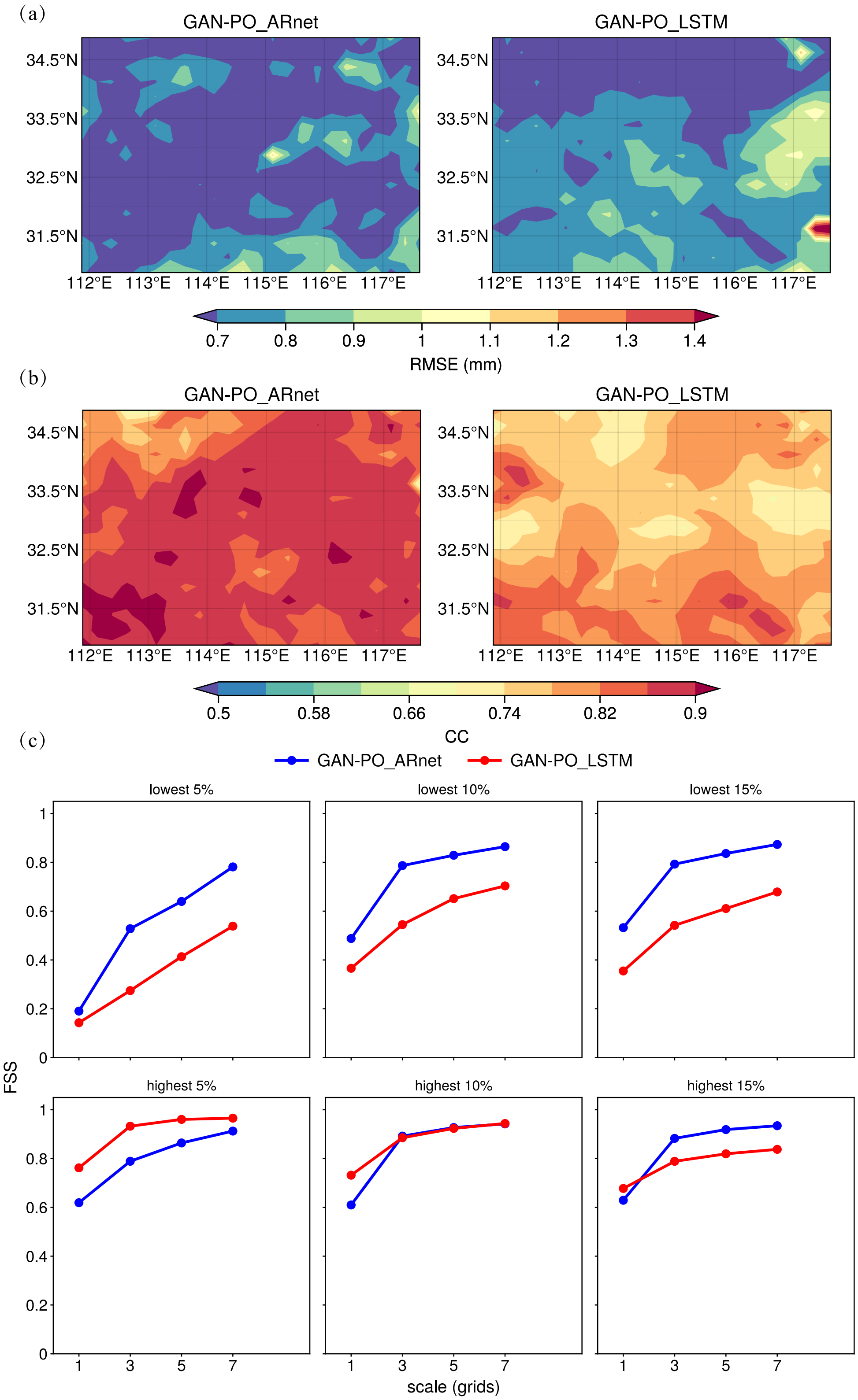

Figure 10. Evaluation results of the simulated ET using parameters calibrated by the GAN-PO method with two different surrogate models. (a) Spatial map of RMSE, (b) spatial map of CC and (c) FSS.

Table 2
Comparison of Computational Time for Different Methods

Method	Training time per epoch (min)	Total computational time (min)
SCE-UA	–	705
dPL	0.90	180
GAN-PO	0.76	152

observed that the L1 and L_R losses decrease with the epoch number, whereas the adversarial loss increases. Consequently, their total sum continues to decrease, indicating that the training process is effective. Even though the L1 loss is relatively large compared to the generator's adversarial loss, reducing λ_l to smaller values does not lead to improved model performance. Some previous studies have also utilized relatively large weights for the L1 regularization term (Isola et al., 2017; Ravuri et al., 2021; Zhu, Park, et al., 2017). As GANs are known for their difficulty in convergence, there are instances where the generator loss may increase even if the quality of the generated images improves. Therefore, it is crucial to utilize evaluation metrics rather than solely relying on loss values to gauge GAN performance.

In the results section, we compare calibration results based on the outcomes derived from the surrogate models, following the practice employed by Tsai et al. (2021). The DL-based surrogate is programmatically differentiable, serving as the foundational technique for both the dPL and GAN-PO methods. The advantages of differentiable modeling have been thoroughly elucidated by Shen et al. (2023). Once trained, applying the inferred parameters from either dPL or GAN-PO to the surrogate model requires mere seconds, offering significant efficiency advantages over the original process-based model. This is especially beneficial for large-scale flood and drought prediction tasks, as well as climate change impact studies. We also evaluate the performance of the actual VIC model when employing the estimated parameters obtained from various methods. Figure S4 in Supporting Information S1 depicts a decrease in the accuracy of VIC-simulated ET for all three methods compared to that of surrogate model-simulated ET, as the surrogate model cannot perfectly reproduce the behavior of the original model. However, the overall ranking of the three methods remains unchanged. GAN-PO maintains its superiority, demonstrating the ability to reduce the RMSE of simulated ET for 75% of the total grid cells during the testing period, while the corresponding values for dPL and SCE-UA are 71% and 65%, respectively. As indicated by Tsai et al. (2021) and Shen et al. (2023), a numerical model can be reimplemented using DL platforms such as PyTorch or Julia to transform it into a differentiable model. Hence, we intend to explore our method on differentiable distributed hydrological models in future studies.

The computational efficiency of an algorithm is a crucial aspect to consider. Table 2 provides a comparative analysis of the running time required for three distinct models utilized in the study. For the SCE-UA method, the total computational time expended amounts to 705 min. The dPL method necessitates 0.9 min per epoch for training, culminating in a total computational time of 180 min for 200 epochs. In contrast, the GAN-PO method exhibits a training time per epoch of 0.76 min, leading to a total computational time of 152 min for 200 epochs. The extended training time per epoch for the dPL method compared to GAN-PO can be attributed to its utilization of paired data from each grid for training, resulting in a significantly larger number of training samples. In this study, we standardize the number of training epochs for both the dPL and GAN-PO methods to ensure a fair comparison. However, if the early stopping technique is implemented with dPL, the effective running epochs could be condensed to 60, resulting in a total computational time of 55 min. Given that the study area is relatively small, the image size is limited. Consequently, the training time for GAN-PO is expected to increase substantially for larger areas due to the escalating requirement for training samples. We argue a similar scenario applies to dPL, especially considering the substantial increase in grid cells.

While the proposed GAN-PO method demonstrates promising results, there is still room for further improvement. The generator in GAN-PO learns a mapping from static geophysical attributes to parameter fields, yet it only generates deterministic outputs. The inherent uncertainty in model parameters should be accounted for through parameter distribution estimation. Therefore, the generator should learn a probabilistic mapping to enable the sampling of various parameter sets, thereby generating simulation results which can capture the predictive distribution. Traditional probabilistic parameter optimization typically relies on Bayesian methods such as Markov Chain Monte Carlo approaches. However, applying these methods to distributed parameter estimation can pose challenges due to the significant computational burden and the difficulties in constructing an appropriate likelihood function (Sun et al., 2017). There have been studies that incorporate random latent vectors into the GAN framework, enabling it to produce probabilistic results that are both diverse and reasonable (Laloy et al., 2018; Ravuri et al., 2021; Zhu, Zhang, et al., 2017) in other research areas. Similar methodologies can be explored in our future research endeavors to account for parameter uncertainty within the GAN-PO framework. In this scenario, the generator takes random noise and static geophysical attributes as inputs to produce probabilistic distributed parameter fields. These fields can then be utilized to generate multiple realistic outputs, thereby providing uncertainty estimates.

In this study, we choose ET as the calibration target primarily due to the availability of the long-term consistent GLEAM V3.7 ET product, which serves as a reliable reference in our study area. It should be noted that GAN-PO can be readily applied to other target variables without requiring modifications to the network architecture. To facilitate comparison with the classical MPR-based distributed calibration method and the state-of-the-art dPL method, we only apply the GAN-PO method to a single-objective calibration problem. However, many previous studies have demonstrated that incorporation of various variables enhances calibration robustness by leveraging multiple observations to constrain model parameters (Dembélé, Hrachowitz, et al., 2020; Gong et al., 2015; Liu et al., 2022; Mei et al., 2023). Our future work plans to adapt the GAN-PO method to harness information from multiple observations. To effectively adapt GAN-PO for multi-objective applications, one possible solution is to combine the power of GANs with multi-task learning. Similar strategies have been explored in the machine learning literature (Bai et al., 2018; Liu et al., 2018). Multi-task GANs can effectively handle complex scenarios where multiple factors influence image generation, making them potentially suited for solving the multi-objective optimization problem.

The current GAN-PO method is applied to calibrate the VIC model over a regular sized basin with a spatial resolution of 0.25°. Although the study area is relatively small in terms of image size (24 × 17), we believe the GAN-PO method remains suitable for distributed model calibration over larger areas with medium model resolution. In the field of image processing, GAN models are frequently trained on 256 × 256 images (Isola et al., 2017; Zhu, Zhang, et al., 2017), which corresponds to a scale that approximates a continental area with a spatial resolution of 0.25°. Therefore, there are no technical barriers to applying GAN-PO to large regions of a size similar to 256 × 256 pixels. However, it's worth noting that the required sample size for training will increase accordingly. As the model resolution increases, the corresponding image size within the same spatial extent will also increase accordingly. In our future work, we would like to further test the applicability of the GAN-PO method to distributed modeling in large domains with higher resolutions. In the machine learning literature, numerous studies have, in fact, investigated the feasibility of scaling adversarial learning to large models and data sets (Brock et al., 2018; Karras et al., 2017, 2019). Comparable concepts can be used to help improve our method.

6. Conclusions

Calibration of distributed land surface hydrologic models is a long-standing challenge. The conversion of rich spatiotemporal information contained in both observation and simulation into single performance metrics for calibration often results in models producing uneven spatial performance. In this study, we propose a novel method to address this challenge using the GAN framework. When employing a deep neural network for distinguishing between observation and simulation, the judgments derived from the network could roughly provide insights into the location of model bias. These insights can be used inversely to train a generative network, aiming to generate seamless parameter fields which make simulations less distinguishable from observations. We apply this GAN-PO method for estimating parameters of the VIC model over the Huaihe basin in China. Compared to the conventional calibration approach and the recently proposed DL-based parameter learning method dPL, our method shows a slightly better capability than dPL in reducing model errors at the grid scale. Both methods exhibit a substantial improvement in simulated ET compared to simulations conducted with default parameters. Their performance surpasses that of the conventional calibration approach based on the MPR technique, in which the single RMSE aggregated over space and time is adopted as the objective function. Furthermore, results show that our method outperforms others in maintaining spatial consistency, as evidenced by the spatial evaluations. Notably, the optimization performance of GAN-PO can be further enhanced by employing a suitable surrogate model.

Data Availability Statement

The VIC model version 5.1.0 used for parameter calibration is preserved at https://github.com/UW-Hydro/VIC/tree/5.1.0. The CMFD data set is available via https://doi.org/10.11888/AtmosphericPhysics.tpe.249369.file. The GLEAM version 3.7 ET product is available via https://www.gleam.eu/. The China Data set of Soil Hydraulic

Parameters and the Soil Database of China for Land Surface Modeling are available via http://globalchange.bnu.edu.cn/research/data. Our whole workflow of parameter learning was implemented in PyTorch version 1.12.1. The source code and the input data for the GAN-PO method are accessible via https://doi.org/10.5281/zenodo.11123631 (Sun et al., 2024).

Acknowledgments

This study is jointly supported by the Second Tibetan Plateau Scientific Expedition and Research Program (20190ZKK0208), the National Natural Science Foundation of China (42101046), the National Key R&D Program of China (2021YFC3201102, 2023YFC3007705), the Key Scientific and Technological Project of the Ministry of Water Resources. P.R.C (SKS-2022001), the Fundamental Research Funds for the Central Universities (B240201161), and the Scientific Research Start-up Foundation for High-level Talents of Hohai University (522020012).

References

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*(12), 10078–10091. https://doi.org/10.1002/2015wr017498

Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 206–221).

Beck, H. E., Pan, M., Lin, P., Seibert, J., Dijk, A. I. J. M., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17). https://doi.org/10.1029/2019jd031485

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., et al. (2021). Advances in land surface modelling. *Current Climate Change Reports*, *7*(2), 45–71. https://doi.org/10.1007/s40641-021-00171-5

Brock, A., Donahue, J., & Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., et al. (2017). The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, *21*(7), 3427–3440. https://doi.org/10.5194/hess-21-3427-2017

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*(3), 2350–2365. https://doi.org/10.1002/2015WR017910

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., et al. (2021). The abuse of popular performance metrics in hydrologic modeling. *Water Resources Research*, *57*(9). https://doi.org/10.1029/2020wr029001

Dai, Y., Shangguan, W., Duan, Q., Liu, B., Fu, S., & Niu, G. (2013). Development of a China dataset of soil hydraulic parameters using pedotransfer functions for land surface modeling. *Journal of Hydrometeorology*, *14*(3), 869–887. https://doi.org/10.1175/jhm-d-12-0149.1

Dembélé, M., Ceperley, N., Zwart, S. J., Salvadore, E., Mariethoz, G., & Schaefli, B. (2020a). Potential of satellite and reanalysis evaporation datasets for hydrological modelling under various model calibration strategies. *Advances in Water Resources*, *143*, 103667. https://doi.org/10.1016/j.advwatres.2020.103667

Dembélé, M., Hrachowitz, M., Savenije, H. H. G., Mariéthoz, G., & Schaefli, B. (2020b). Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets. *Water Resources Research*, *56*(1). https://doi.org/10.1029/2019wr026085

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*(4), 1015–1031. https://doi.org/10.1029/91WR02985

Feigl, M., Herrnegger, M., Klotz, D., & Schulz, K. (2020). Function space optimization: A symbolic regression method for estimating parameter transfer functions for hydrological models. *Water Resources Research*, *56*(10), e2020WR027385. https://doi.org/10.1029/2020WR027385

Feigl, M., Thober, S., Schweppe, R., Herrnegger, M., Samaniego, L., & Schulz, K. (2022). Automatic regionalization of model parameters for hydrological models. *Water Resources Research*, *58*(12), e2022WR031966. https://doi.org/10.1029/2022WR031966

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., et al. (2015). Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences*, *19*(5), 2409–2425. https://doi.org/10.5194/hess-19-2409-2015

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Gou, J., Miao, C., Samaniego, L., Xiao, M., Wu, J., & Guo, X. (2021). CNRD v1.0: A high-quality natural runoff dataset for hydrological and climate studies in China. *Bulletin of the American Meteorological Society*, *102*(5), E929–E947. https://doi.org/10.1175/bams-d-20-0094.1

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., & Mao, Y. (2018). The variable infiltration capacity model version 5 (VIC-5): Infrastructure improvements for new applications and reproducibility. *Geoscientific Model Development*, *11*(8), 3481–3496. https://doi.org/10.5194/gmd-11-3481-2018

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing dataset for land process studies over China. *Scientific Data*, *7*(1), 25. https://doi.org/10.1038/s41597-020-0369-y

Hundecha, Y., & Bárdossy, A. (2004). Modeling of the effect of land use changes on the runoff generation of a river basin through parameter regionalization of a watershed model. *Journal of Hydrology*, *292*(1), 281–295. https://doi.org/10.1016/j.jhydrol.2004.01.002

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Jia, Y., Li, C., Yang, H., Yang, W., & Liu, Z. (2022). Assessments of three evapotranspiration products over China using extended triple collocation and water balance methods. *Journal of Hydrology*, *614*, 128594. https://doi.org/10.1016/j.jhydrol.2022.128594

Jiang, L., Wu, H., Tao, J., Kimball, J. S., Alfieri, L., & Chen, X. (2020). Satellite-based evapotranspiration in hydrological model calibration. *Remote Sensing*, *12*(3), 428. https://doi.org/10.3390/rs12030428

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004362

Koch, J., Jensen, K. H., & Stisen, S. (2015). Toward a true spatial model evaluation in distributed hydrological modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and evaluated against a modeling case study. *Water Resources Research*, *51*(2), 1225–1246. https://doi.org/10.1002/2014wr016607

Laloy, E., Hérault, R., Jacques, D., & Linde, N. (2018). Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resources Research*, *54*(1), 381–406. https://doi.org/10.1002/2017wr022148

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483

Liu, X., Yang, K., Ferreira, V. G., & Bai, P. (2022). Hydrologic model calibration with remote sensing data products in global large basins. *Water Resources Research*, *58*(12), e2022WR032929. https://doi.org/10.1029/2022WR032929

Liu, Y., Wang, Z., Jin, H., & Wassell, I. (2018). Multi-task adversarial network for disentangled feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3743–3751).

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *Journal of Geophysical Research*, *109*(D7). https://doi.org/10.1029/2003JD003517

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2794–2802).

Mao, Y., Bai, J., Wu, G., Xu, L., Yin, C., Feng, F., et al. (2024). Terrestrial evapotranspiration over China from 1982 to 2020: Consistency of multiple data sets and impact of input data. *Journal of Geophysical Research: Atmospheres*, *129*(3), e2023JD039387. https://doi.org/10.1029/2023JD039387

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., et al. (2017). GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, *10*(5), 1903–1925. https://doi.org/10.5194/gmd-10-1903-2017

Mei, Y., Mai, J., Do, H. X., Gronewold, A., Reeves, H., Eberts, S., et al. (2023). Can hydrological models benefit from using global soil moisture, evapotranspiration, and runoff products as calibration targets? *Water Resources Research*, *59*(2), e2022WR032064. https://doi.org/10.1029/2022WR032064

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017wr020401

Oubeidillah, A. A., Kao, S. C., Ashfaq, M., Naz, B. S., & Tootle, G. (2014). A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US. *Hydrology and Earth System Sciences*, *18*(1), 67–84. https://doi.org/10.5194/hess-18-67-2014

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., Lee, J., et al. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509. https://doi.org/10.1029/2021ms002509

Pokhrel, P., & Gupta, H. V. (2010). On the use of spatial regularization strategies to improve calibration of distributed watershed models. *Water Resources Research*, *46*(1). https://doi.org/10.1029/2009wr008066

Pokhrel, P., Gupta, H. V., & Wagener, T. (2008). A spatial regularization approach to parameter estimation for a distributed watershed model. *Water Resources Research*, *44*(12). https://doi.org/10.1029/2007WR006615

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Refsgaard, J. C. (1997). Parameterisation, calibration and validation of distributed hydrological models. *Journal of Hydrology*, *198*(1), 69–97. https://doi.org/10.1016/S0022-1694(96)03329-X

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, Proceedings, Part III*, *18*, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008wr007327

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., et al. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, *21*(9), 4323–4346. https://doi.org/10.5194/hess-21-4323-2017

Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L., et al. (2013). A China data set of soil properties for land surface modeling. *Journal of Advances in Modeling Earth Systems*, *5*(2), 212–224. https://doi.org/10.1002/jame.20026

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth and Environment*, *4*(8), 552–567. https://doi.org/10.1038/s43017-023-00450-9

Sun, R., Duan, Q., & Huo, X. (2021). Multi-objective adaptive surrogate modeling-based optimization for distributed environmental models based on grid sampling. *Water Resources Research*, *57*(11), e2020WR028740. https://doi.org/10.1029/2020WR028740

Sun, R., Pan, B., & Duan, Q. (2023). A surrogate modeling method for distributed land surface hydrological models based on deep learning. *Journal of Hydrology*, *624*, 129944. https://doi.org/10.1016/j.jhydrol.2023.129944

Sun, R., Pan, B., & Duan, Q. (2024). Generative adversarial network-based parameter optimization (GAN-PO). *Zenodo*. https://doi.org/10.5281/zenodo.11123631

Sun, R., Yuan, H., & Liu, X. (2017). Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty estimation. *Journal of Hydrology*, *554*, 680–692. https://doi.org/10.1016/j.jhydrol.2017.09.041

Troy, T. J., Wood, E. F., & Sheffield, J. (2008). An efficient calibration method for continental-scale land surface modeling. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007wr006513

Tsai, W. P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., et al. (2021). From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature Communications*, *12*(1), 5988. https://doi.org/10.1038/s41467-021-26107-z

Xia, Y., Mocko, D. M., Wang, S., Pan, M., Kumar, S. V., Peters-Lidard, C. D., et al. (2018). Comprehensive evaluation of the variable infiltration capacity (VIC) model in the north American land data assimilation system. *Journal of Hydrometeorology*, *19*(11), 1853–1879. https://doi.org/10.1175/JHM-D-18-0139.1

Yang, Y., Pan, M., Beck, H. E., Fisher, C. K., Beighley, R. E., Kao, S. C., et al. (2019). Quest of calibration density and consistency in hydrologic modeling: Distributed parameter calibration against streamflow characteristics. *Water Resources Research*, *55*(9), 7784–7803. https://doi.org/10.1029/2018wr024178

Yu, S., & Ma, J. (2021). Deep learning for geophysics: Current and future trends. *Reviews of Geophysics*, *59*(3). https://doi.org/10.1029/2021rg000742

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*, (pp. 2223–2232).

Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017). Toward multimodal image-to-image translation. *Advances in Neural Information Processing Systems*, *30*.