On numerical methods and differentiable modeling for soil process representations in the NextGen Framework in arid regions

Ryoko Araki[1], Soelem Aafnan Bhuiyan[2], Tadd Bindas[3], Jeremy Rapp[4]
*All authors contributed equally, and order is set alphabetically

[1]San Diego State University & University of California, Santa Barbara, California; *raraki8159@sdsu.edu*
[2]George Mason University, Fairfax, Virginia; *sbhuiya2@gmu.edu*
[3]Penn State University, University Park, Pennsylvania; *tkb5476@psu.edu*
[4]Michigan State University, East Lansing, Michigan; *rappjer1@msu.edu*

Academic Advisors: Hilary McMillan, *San Diego State University*; Viviana Maggioni, *George Mason University*; Chaopeng Shen, *Penn State University*; Anthony Kendall, *Michigan State University*
Summer Institute Theme Advisors: Jonathan M. Frame, Floodbase, *jonathan@floodbase.com*

Abstract: Accurate streamflow predictions are integral in managing and forecasting riverine flooding. Previous National Water Center (NWC) Summer Institute teams have explored the evaluation and ensembling of multiple hydrologic streamflow prediction models over CONUS. Their multi-model approach shows high Normalized Nash-Sutcliffe Efficiency (NNSE) performances in most basins but underperforms in arid/semiarid regions. We hypothesize soil/physical processes in these regions are not adequately represented within streamflow prediction models. Thus, there is a need to create a dynamic framework capable of making better predictions while more accurately considering soil/physical processes represented in the system. We developed an Ordinary Differential Equation (ODE) representation of soil fluxes within the Conceptual Functional Equivalent Framework (CFE) model and compared it to the existing framework using observed streamflow from 498 (Catchment Attributes and Meteorology for Large-sample Studies) CAMELS basins. Results were inconclusive as CFE with the ODE soil flux integration performed better than CFE according to NSE but worse regarding Kling-Gupta Efficiency (KGE). To further investigate the representation of soil/physical properties in arid/semiarid regions, we created differential programming approaches for CFE and the Layered Green & Ampt with Redistribution (LGAR) model to determine physically-informed model parameters. All project deliverables are available inside an open-source GitHub repository.

---

1. Motivation

In recent years, a major focus of hydrologic research has been on evaluating the performance of models for streamflow forecasting within the contiguous United States (CONUS) over a benchmark dataset (CAMELS; (Addor et al., 2017; Newman, Andrew, 2014)). These efforts have compared the performance of physical-conceptual-based models against emergent deep learning models across various catchment scales and hydrologic settings (Feng et al., 2020; Kratzert et al., 2019b). Within these analyses, a ubiquitous performance drop in catchments under arid/semi-arid climates, where soil processes greatly control hydrologic connectivity (Araki et al., 2022, 2023), has been observed for all model schema (Feng et al., 2020; Lahmers et al., 2021). Along with the core physical processes in the arid regions, key characteristics such as phase correlation, forest, and grass cover restrict model capacity to accurately estimate regional evapotranspiration in arid regions (Johnson et al., 2023). Hence, it is evident that models need a better representation of the land-surface or soil module to close the balance, which has already been posited by Keith Beven in his 2023 paper of "what models and parameter sets might be considered as not fit-for-purpose (Beven, 2023)".

Also in line with Beven 2023, we attempted to establish "a common framework for assessing model performance to allow consideration of data uncertainty" by evaluating model performance in arid/semiarid regions within the NextGen framework. To this end of hypothesis testing, multiple fit-for-purpose parameter settings and models are the key requirements in the common evaluation framework outlined in (Beven, 2023). Of the models concerning soil processes, "may have to add the citation name here"(Johnson et al., 2023) have explored five evapotranspiration models to better represent potential evapotranspiration to close the water balance and to limit the influx going to soil modules and, therefore, its runoff-threshold processes. However, soil moisture models are yet to be tested in such a framework. Spatial heterogeneity of soil parameters plays a more crucial role in estimating catchment-scale runoff thresholds in arid/semiarid regions (Western et al., 2001). As there exists no single best global hydrologic model, regionalized hydrologic modeling via a multi-model approach is preferred. The next-gen framework includes a range of conceptual and deep-learning models such as CFE, LSTM, and TOPMODEL. Among these models, CFE has a more dedicated module on the soil moisture-runoff relationship. Combining multiple models through differentiable modeling (DM) in geosciences (Shen et al., 2023) is an opportunity to better examine internal physical processes to further validate the Beven hypothesis.

Using a multiple-model approach in the next-gen framework, we test the null hypothesis "models with detailed soil process representation perform no better than the ones with less detailed soil in the arid/semiarid regions" in CAMELS basins. Our multimodel approach consists of a combination of process-based and differentiable models considered to be "fit-for-the purpose" (Beven, 2023) in arid regions, i.e., models with finer soil process representation empowered with a parameter-learning system for predicting hydrograph patterns (See model description in Section 4.3 and 4.4). First, we introduce the ordinary differential equation (ODE) to Conceptual Functional Equivalent (CFE) model (Section 4.3.2); this change allows a better representation of the interaction between multiple outflows from soil storage, preserving the original CFE principles. Additionally, we selected the Layered Green & Ampt with Redistribution model (LGAR) (Section 4.3.4), a process-based model that mimics the Richards/Richardson equation to partition infiltration/runoff meant to predict simulations in arid environments. To further study the integration of soil processes and arid/semiarid runoff, we introduce differentiable versions of the CFE, CFE-ODE, and LGAR methods (Section 4.4). Regarding catchment and parameter configuration, previous benchmarking studies have generated model outputs at the CAMELS catchment extent in lumped configurations. To improve the representation of soil spatial heterogeneity, we worked on extracting CAMELS benchmarking datasets at a subcatchment scale (Hydrofabric artifact scale, 4-10 sq km) using model parameters calibrated at the lumped catchment scale (See Conclusion).

2. Objective and Scope

We developed process-based and differentiable models with detailed representation of soil processes, and evaluated their performance in arid environments. Our research questions are:

1. Can we improve model predictions from CFE in arid regions by updating the soil moisture partitioning using an ordinary differential equation (ODE)? (See the results in Section 5.1).
2. Can differentiable modeling show an improvement over process-based models in arid regions or assist in learning intermediate soil/physical processes within CFE and LGAR? (See the results in Section 5.2).

3. Previous Studies

Our work uses the NOAA-OWP NWC NextGen framework and builds upon the project by the 2022 Summer Institute team, “Automated decision support for model selection in the Nextgen National Water Model" (Deardorff et al., 2022). While their study focused on model selection over CONUS, we focused on investigating multi-model approaches in arid regions, specifically using models with high representations of soil properties. We generated an open-source [GitHub repository] that enables researchers and developers to reproduce and further enhance our work.

4. Methodology

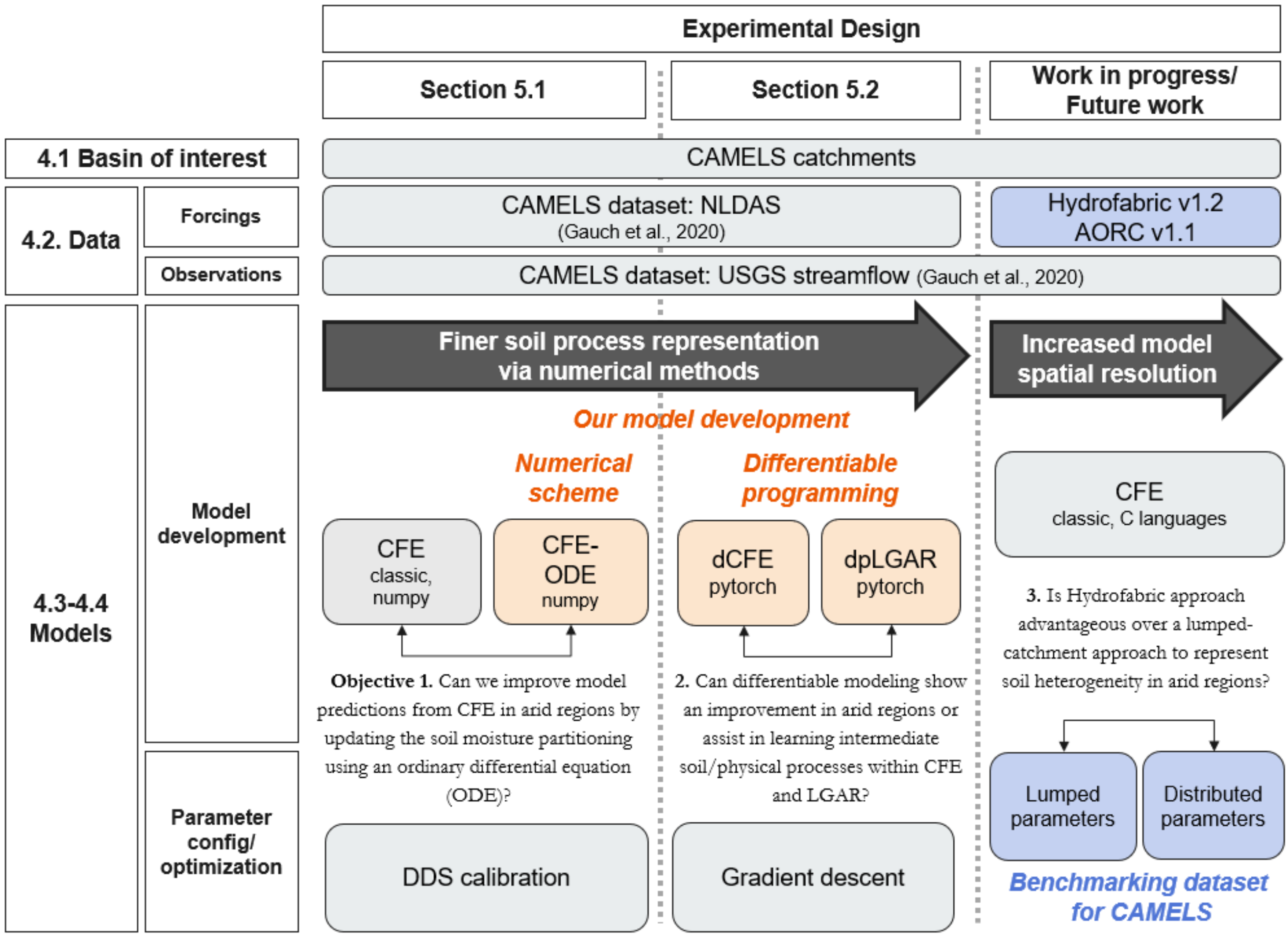


Figure 1: Overview of our experimental design, with the corresponding Method section on the left and Result section at the top. Model comparisons and corresponding research objectives/questions are shown in the middle.

4.1. Basins of interest

The Catchment Attributes and Meteorological dataset (CAMELS) is a hydrologic dataset that allows large-sample studies, with catchments minimally impacted by human activities (Addor et al., 2017; Newman, Andrew, 2014). To ensure consistency, we used 498 CAMELS catchments for continuous United States (CAMELS-US) used in previous benchmarking studies (Gauch et al., 2021; Kratzert et al., 2019a; Newman et al., 2017). Out of 617 CAMELS-US basins, 516 catchments are chosen following criteria by (Gauch et al., 2021). First, (Gauch et al., 2020) excluded basins with high uncertainty in the definition of basin area, where a large (>10%) discrepancy was identified in basin area depending on the calculation method (Kratzert et al., 2019a; Newman et al., 2017). Second, they excluded basins larger than 2,000 km^2 (Kratzert et al., 2019a; Newman et al., 2017). Last, they excluded basins where USGS hourly stream gauge data were not available (Gauch et al., 2021). We further reduced the number of basins to 498 during model calibration. Due to a lack of hourly streamflow observation data for the model calibration period (See Section 4.3.3. for the details), 10 basins were left out in the calibration process. In addition, 2 basins failed to converge during the calibration runs.

4.2. The CAMELS dataset: Forcing and observed data:
Following the previous benchmarking studies, we used a dataset developed by Gauch et al. (2020, 2021), which includes hourly streamflow data from the United States Geological Survey (USGS) and forcing data (hourly total rainfall and potential evapotranspiration) from the National Land Data Assimilation System (NLDAS) for CAMELS basin. Specifically for the model evaluation in the arid catchment using Hydrofabric, we chose statistical metrics such as Kling-Gupta Efficiency (KGE) and Nash-Sutcliffe Efficiency (NSE). We have also applied the Cumulative Distribution Function (CDF) to get insights into the distribution of the statistical metrics.

4.3. Models
4.3.1. CFE model (https://github.com/NWC-CUAHSI-Summer-Institute/cfe_py)
Conceptual Functional Equivalent (CFE) is a conceptual hydrologic model that provides similar functionality to the current National Water Model with simplified soil moisture and routing expressions (Ogden, n.d.). The current National Water Model based on the WRF-Hydro uses the Clapp-Hornberger relation (Clapp & Hornberger, 1978) for the water retention function (a relationship between the soil suction and soil water content). And when the soil water content exceeds field capacity, lateral runoff is generated. The CFE model redefines this threshold for generating lateral flow and percolation by using the hydrostatic equilibrium based on soil suction instead of field capacity.

4.3.2. CFE-ODE model (https://github.com/NWC-CUAHSI-Summer-Institute/cfe_py)
We modified the soil moisture module in CFE. The current version of CFE models sequentially subtracts three fluxes going out from soil moisture storage (soil evaporation, percolation, then lateral flow). This reduces computational time but (1) introduces approximation errors because the fluxes are assumed constant over the timestep and, therefore, (2) inaccurately represents the internal feedback (Clark & Kavetski, 2010). To solve those issues, we introduced an alternative numerical scheme, an ordinary differential equation (ODE), to the soil moisture module in the CFE written in BMI-Python format. The physics equations in the ODE scheme are equivalent to the existing CFE formulation, but they are coded as ordinary differential equations, and the outfluxes are simultaneously calculated and subtracted from the soil moisture storage. Hereafter, we refer to the model with the previous soil moisture calculation scheme (as introduced in 4.3.1.) as "CFE-classic" and with the ODE scheme as "CFE-ODE". In the parameter setting, users are allowed to choose options to run either CFE-classic or CFE-ODE.

4.3.3. CFE calibration (https://github.com/NWC-CUAHSI-Summer-Institute/calibrate-cfe)
Both CFE-classic and CFE-ODE were calibrated for each CAMELS basin. We calibrated the nine parameters and two runoff subroutines (Schaake and Xinanjiang). The calibration parameters are selected, and their bounds are taken from the current National Water Model calibration, with the maximum and minimum parameter values defined in soilproperties_CONUS_FullRouting.nc and GWBUCKPARAM_CONUS_FullRouting.nc (See Table S1). The optimal parameter sets were explored using the Dynamically Dimensioned Search optimization algorithm (DDS) with 500 evaluation runs (Huard, 2023; Tolson & Shoemaker, 2007) using the Kling-Gupta Efficiency as an objective function (Gupta et al.,

2009). We used a Python package SPOTPY (Houska et al., 2015) for calibration. The calibration and testing periods were the same as the LSTM training period of WY 2000-2008 (October 1, 1999 - September 30, 2008) and the testing period of WY 2010 (October 1, 2009 - September 30, 2010), respectively. A one-year spin-up period was implemented prior to each calibration and validation period.

[4.3.4. LGAR model (https://github.com/NOAA-OWP/LGAR-C)](https://github.com/NOAA-OWP/LGAR-C)

The Layered Green & Ampt with Redistribution (LGAR) (La Follette et al., 2023) model partitions precipitation into infiltration and runoff and was designed to mimic the Richardson-Richards (RRE) equation (Richards, 1931; Richardson, 1922) for semi-arid and arid basins. LGAR can be conceptually visualized as a bucket filled with different layers of soil (Figure 2a). Each layer possesses different Van Genuchten (van Genuchten, 1980) parameters and hydrologic conductivity values (K_{sat}), which control water storage. Through tracking water storage, soil infiltration, evapotranspiration, runoff, and percolation can be inferred. Van Genuchten parameters, and hydrologic conductivity values, are statically defined for each soil column via a soil survey to a lookup table. Figure 2b shows a simple example of volumetric water content and pressure head, varying with depth.

(a) (b)

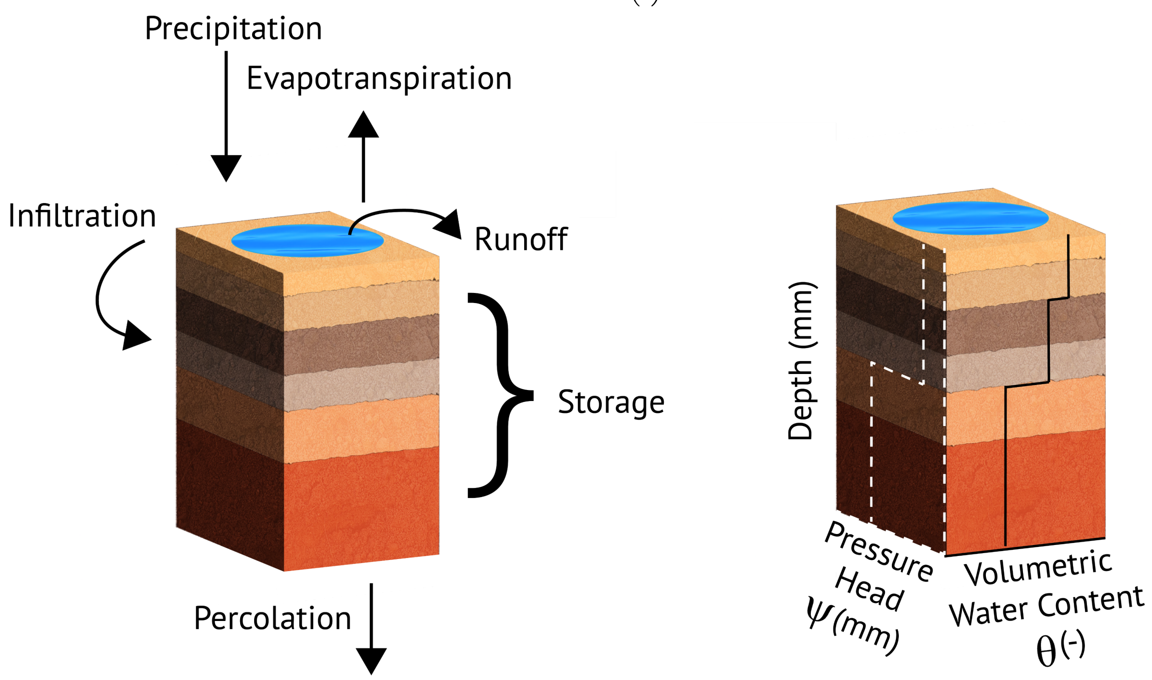


Figure 2: (a) A conceptual representation of the hydrologic properties within a soil column. (b) Two graphs show how pressure head and volumetric water content change through soil layers.

4.4. Differentiable modeling

[4.4.1. dpLGAR (https://github.com/NWC-CUAHSI-Summer-Institute/dpLGAR)](https://github.com/NWC-CUAHSI-Summer-Institute/dpLGAR)

By implementing LGAR on a differentiable platform (PyTorch), we can train a NN to produce physical soil parameters representative of each soil layer, similar to work done in previous differentiable parameter learning (dPL) studies (Bindas et al., 2023; Feng et al., 2022, 2023; Tsai et al., 2021). We input *p* number of SSURGO (POLARIS) Polaris (Chaney et al., 2019) two-meter soil column attributes into a Multilayer Perceptron Model (MLP) (Leshno et al.,

1993) (Figure 3) to learn lumped catchment scale soil parameters and ponding depth limitations.

$$\alpha, n, K_{sat}, \theta_s, \theta_r, d_p = NN(c) \quad (1)$$

Attributes used for parameter prediction include soil composition percentages (clay, silt, sand), soil pH, and organic matter content (Table S2).

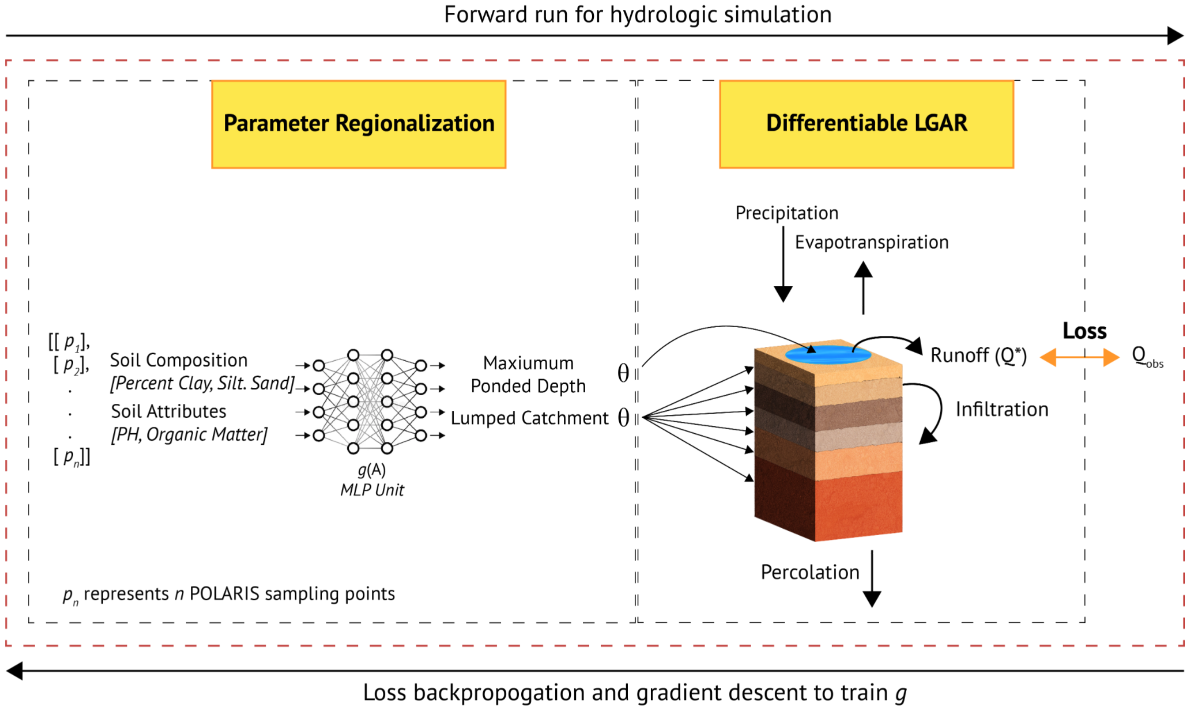


Figure 3: A flowchart showing the integration of LGAR and the parameter-learning NN scheme used to generate lumped catchment soil parameters from *n* POLARIS sampling points

4.4.2. dCFE (https://github.com/NWC-CUAHSI-Summer-Institute/dCFE)

Similar to dpLGAR, we implemented the CFE model (both CFE-classic and CFE-ODE) in a differentiable platform (PyTorch) to learn CFE parameters (Figure 4); the code is available from the above URL. The two parameters learned are refkdt and satdk, both highly important to determine the influx to soil water storage (Zhang et al., 2020). We input four static attributes from the Hydrofabric for one CAMELS basin as a test case into a Multilayer Perceptron Model (MLP) (Leshno et al., 1993) (Table S3).

$$refkdt,\ satdk = NN(c) \quad (2)$$

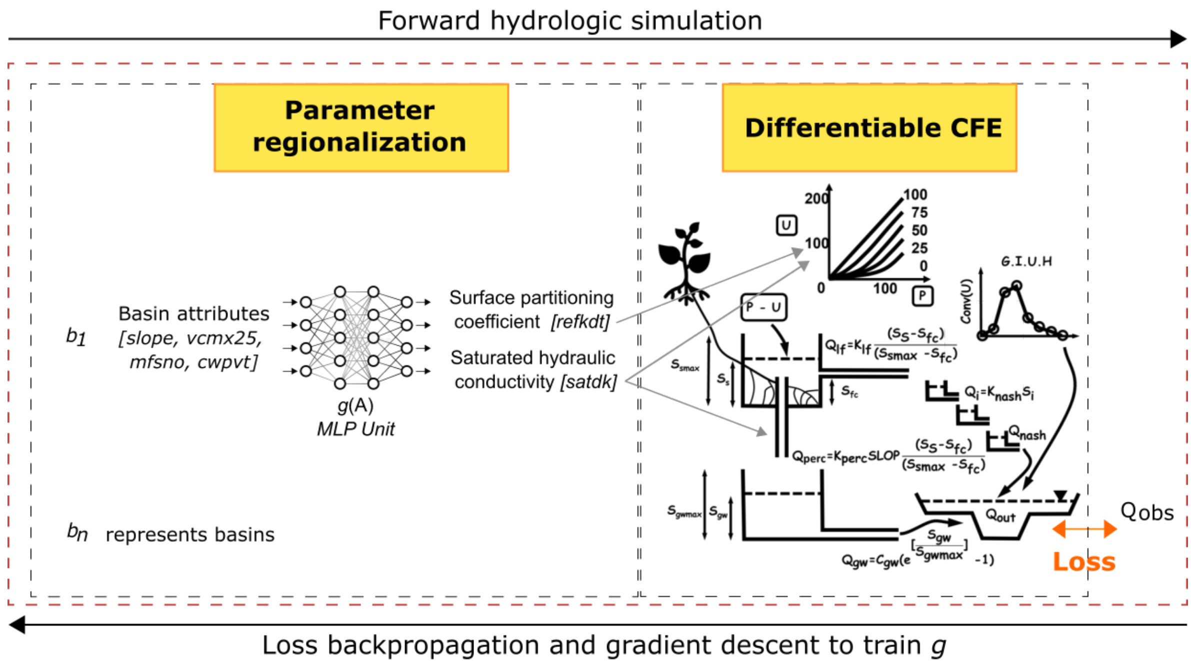


Figure 4: A flowchart showing the integration of CFE and the parameter-learning NN scheme used to generate CFE parameters from a CAMELS basin

4.5. Evaluation

We evaluated the models' streamflow discharge using the following metrics in Table 1.

Table 1. Evaluation metrics used in this study

Metrics	**Description**	**Reference**
NSE	Nash-Sutcliffe Efficiency	Equation (3) in (Nash & Sutcliffe, 1970)
KGE	Kling-Gupta Efficiency	Equation (9) in (Gupta et al., 2009)

We classified aridity of the catchments using the CAMELS Aridity Index.

$$\text{Aridity} = \frac{PET}{P} \tag{3}$$

where it classifies regions with aridity > 1.00 as arid regions. Here, the average precipitation (P) and the potential evapotranspiration (PET) in the CAMELS dataset were calculated from Oak Ridge National Laboratory's (ORNL) Daymet dataset in mm/day unit, starting 1 October 1989 to 30 September 2009 (Addor et al., 2017).

Note that Aridity Index (AI) is more commonly defined as the ratio of precipitation (P) to potential evapotranspiration (PET) (Zomer et al., 2022) as the generalized function of aridity, which is formulated as Equation (2):

$$\text{AI} = \frac{P}{PET} \tag{4}$$

The AI classifies aridity with hyper-arid (AI < 0.05), arid (0.05 <= AI < 0.2), semi-arid (0.2 <= AI < 0.5), dry subhumid (0.5 <= AI < 0.65) and humid (AI => 0.65) regions. To ensure consistency, the analysis and plotting of the results were performed using aridity from the CAMELS dataset.

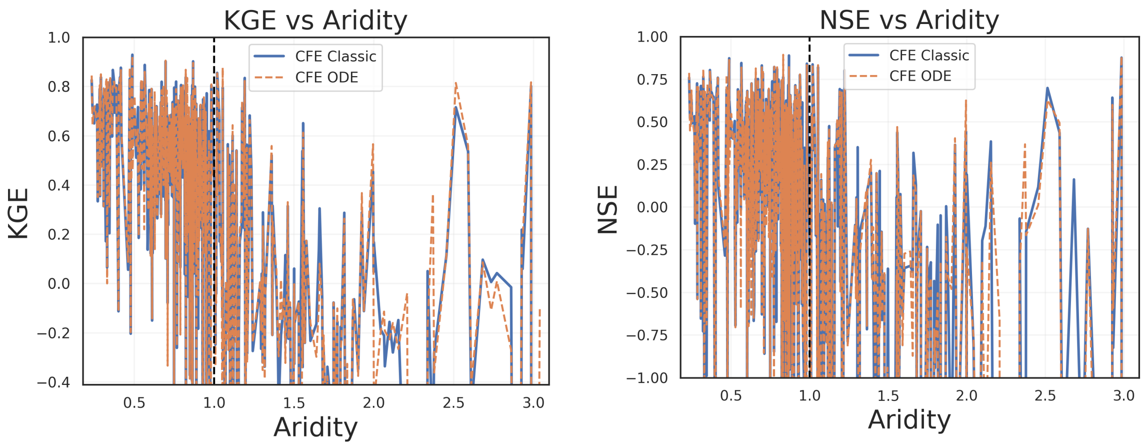
5. Results and Discussion

5.1. Model evaluation between CFE-classic and CFE-ODE in arid regions

Corresponding to the objective/question 1: Can we improve model predictions from CFE by updating the soil moisture partitioning using an ordinary differential equation (ODE)?

In sum, enhancing soil process representation by introducing an ordinary differential equation did not result in improved streamflow prediction in arid regions (see the following subsections for details), and therefore, we pursue even finer soil representation in Section 5.2.

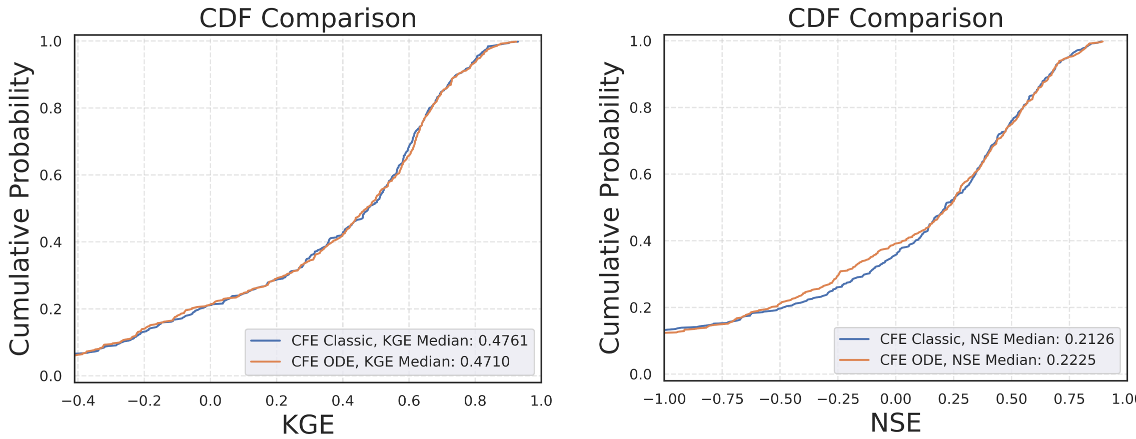
Similar to previous studies, we observed a CFE performance slump in arid regions (Figure 5). The performance in basins with aridity values beyond 1 rarely exceeded the KGE = -0.41 and NSE = 0.0, indicating the simulation was no better than the mean flow (Knoben et al., 2019).



(a) (b)

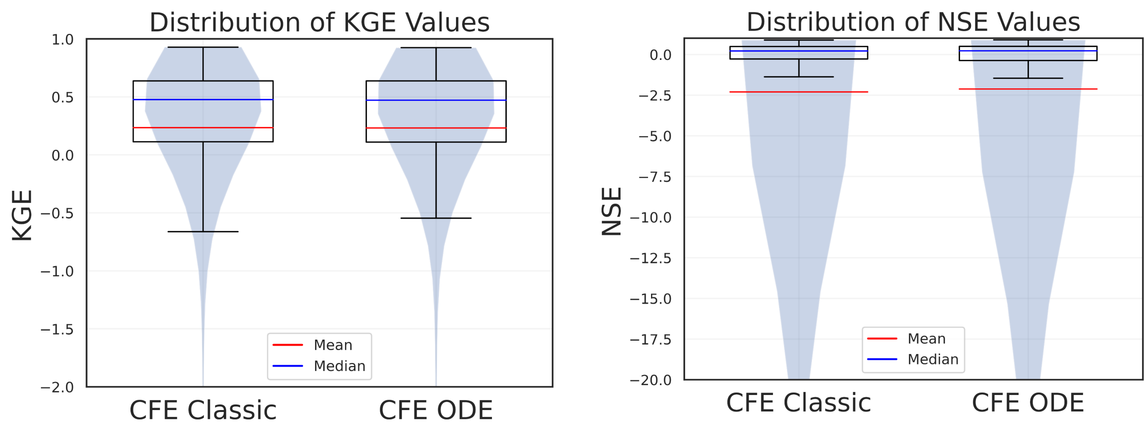
Figure 5: The (a) KGE and (b) NSE performance with increasing aridity for the 498 CAMELS basins. Aridity values beyond 1, on the right-hand side of the vertical black dashed line, are classified as arid regions. Note that extremely low KGE and NGE outliers were beyond the Y-axis limits of the figure.

The performance of the CFE-classic and ODE were similar in terms of KGE and NSE values (Figures 6 & 7). CFE-classic had more basins on the lower end of NSE metrics values in terms of NSE (Figures 6b & 7b). There were no apparent geographic patterns in terms of which model performed better between the CFE-classic or ODE (Figure 8).



(a) (b)

Figure 6: Cumulative distribution of the (a) KGE and (b) NSE on the CFE-classic and CFE-ODE for the 498 CAMELS basins. Note that extremely low KGE and NGE outliers were beyond the X-axis limits of the figure.



(a) (b)

Figure 7: The distribution of (a)KGE and (b)NSE performance for the CAMELS 498 basins. Note that extremely low KGE and NGE outliers were beyond the X-axis limits of the figure.

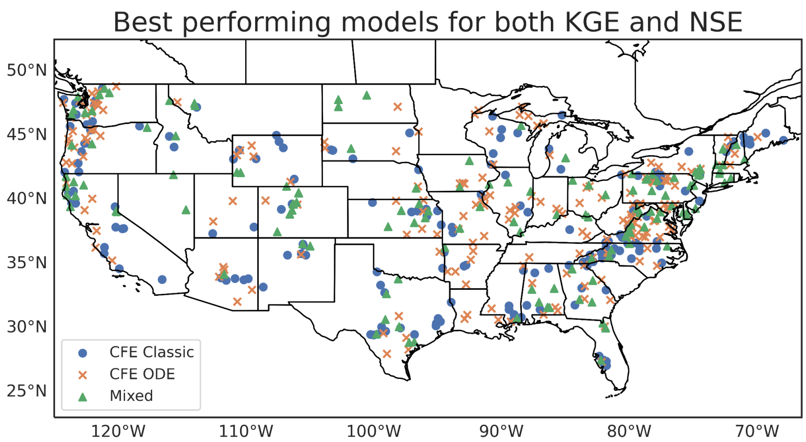


Figure 8: CAMELS basins gauge locations showing the best-performing models between CFE Classic and CFE ODE for both KGE and NSE values. The blue circles show the gauges where both KGE and NSE agree that CFE Classic performed better. The orange crosses show the gauges where both KGE and NSE agree that CFE ODE performed better. The green triangles show the gauges where the best performing models were undetermined; CFE Classic performed better for KGE and CFE ODE performed better for NSE and vice versa.

There are a few possible explanations as to why introducing ODE did not improve the model performance. First, it is possible that parameter calibration compensated for the conceptual differences between CFE-classic and CFE-ODE models. The CFE-classic model is designed to prioritize certain outfluxes to deplete soil moisture storage. However, during the calibration, the parameters may have converged to the values that allowed all outfluxes an equal chance to consume soil moisture storage.

Second, by visually inspecting the hydrograph, we observed a few cases where the observation and models behaved in completely different ways in arid/semiarid regions. The observed streamflow showed no runoff peaks, whereas rainfall and simulation did, indicating that the partitioning between infiltration and runoff did not match up between observation and simulation. Additionally, for a few catchments, the observed streamflow was delayed by months, suggesting that extremely slow baseflow or snowmelt was happening. These errors can be attributed to (1) disinformative data (i.e., errors in the observed streamflow data) or (2) the conceptual design of CFE would struggle to replicate the observed behavior. To be able to distinguish between the possible causes for these errors, additional investigation on water balance residuals is needed to exclude data periods with unreliable datasets (Beven & Westerberg, 2011).

5.2. Demonstrating differentiable versions of LGAR and CFE

Corresponding to the objective/question 2: Can differentiable modeling show an improvement over process-based models in arid regions or assist in learning intermediate soil/physical processes within CFE and LGAR?

5.2.1. dpLGAR (https://github.com/NWC-CUAHSI-Summer-Institute/dpLGAR)

The differentiable LGAR model is still being benchmarked against synthetic soil parameters. Code for the model and the state of the project can be found here [INSERT ZENODO LINK].

5.2.2. dCFE (https://github.com/NWC-CUAHSI-Summer-Institute/dCFE)

The differentiable CFE model is still being benchmarked against synthetic parameters. We plan to expand this to incorporate multiple basins and time-varying attributes. The state of the project can be accessed through the above GitHub link.

We demonstrate the gradient chain in this visualization (see this PDF CFE_gradient_chain_demo.pdf) using PyTorchViz, for a case where the soil moisture storage is active. Every time a parameter (e.g., maximum soil water content) interacts with a flux (e.g., lateral flux, percolation flux), gradients are calculated. These gradient chains aid in tuning parameters via loss backpropagation. This visualization also raises a concern regarding conditional branching in hydrologic modeling. Most of today's conceptual hydrologic models use "if" statements - e.g., "if" precipitation exceeds the infiltration limit, "then" then surface runoff is generated. In the visualization, CFE groundwater parameters (Cgw and expon) are not contributing to the runoff because the conditional statement "if groundwater storage is less than the maximum, exponential nonlinear outflux is calculated" was never triggered. We observed in both dpLGAR and dCFE development that a number of logic gates need to be trigged to generate runoff.

6. Conclusion

Leveraging the Next-gen framework, we tackled the long-standing issue in hydrologic modeling: the substandard predictive capability of streamflow in arid/semiarid regions to refute or confirm the null hypothesis posited in (Beven, 2023). The overall CFE model performance confirmed the findings of previous studies, revealing a rapid performance decline in streamflow NSE and KGE for basins where evapotranspiration (PET) exceeds precipitation (P).

In our pursuit of a more comprehensive representation of soil outflux calculation, we introduced ordinary differential equations (ODEs) to the CFE soil module. However, we found that ODE solutions provided only marginal improvements, indicating simple parameter tuning outweighs the complexity of the model. Further modeling at finer spatial scales using the Open Geospatial Consortium compliant hydrology artifact model application dataset (Hydrofabric) over the CAMELS extents could reveal greater performance differences. While this was originally intended to be conducted during the 2023 Summer Institute timeframe, data access limitations to forcing data (AORC v1.1), software versioning issues (Hydrofabric "pre-release" version versus the current Hydrofabric v1.2), and incomplete software documentation for successful deployment of NextGen pushed this task out of scope.

Furthermore, our project aimed to bridge the recent divide between machine learning models and physics/conceptual-based modeling. DM methods were applied to both CFE and LGAR. However, more work is required in parameter recovery and model tuning.

Supplementary Materials (Data and code availability)

We adhered to a reproducible science protocol utilizing Hydroshare and GitHub repository. You can access our main repository at (https://github.com/NWC-CUAHSI-Summer-Institute/ngen-aridity), which provides a comprehensive summary of all available resources, including our models, dataset, and accompanying pipelines.

Acknowledgments

We'd like to thank CUASHI, the National Water Center at the University of Alabama, and CIROH for funding/support. We'd like to acknowledge and thank Dr. James Halgren, Peter La Follette, and Ahmad Jan for assisting with running LGAR, and Dr. Tony Castronova, Dr. Irene Garousi-Nejad, Dr. Nels Fraizer, Dr. Luciana Cunha, Dr. Zach Wills, Arpita Patel for providing us with valuable advice on running Next-gen and Hydrofabric.

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Araki, R., Branger, F., Wiekenkamp, I., & McMillan, H. (2022). A signature-based approach to quantify soil moisture dynamics under contrasting land-uses. *Hydrological Processes*, *36*(4), e14553. https://doi.org/10.1002/hyp.14553

Araki, R., Mu, Y., & McMillan, H. (2023). Evaluation of GLDAS soil moisture seasonality in arid climates. *Hydrological Sciences Journal*, *0*(ja), null. https://doi.org/10.1080/02626667.2023.2206032

Beven, K. (2023). Benchmarking hydrological models for an uncertain future. *Hydrological Processes*, *37*(5), e14882. https://doi.org/10.1002/hyp.14882

Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, *25*(10), 1676–1680. https://doi.org/10.1002/hyp.7963

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., & Shen, C. (2023). *Improving large-basin river routing using a differentiable Muskingum-Cunge model and physics-informed machine learning* [Preprint]. Preprints. https://doi.org/10.22541/essoar.168500246.67971832/v1

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L. S., McBratney, A. B., Wood, E. F., & Yimam, Y. (2019). POLARIS Soil Properties: 30-m Probabilistic Maps of Soil Properties Over the Contiguous United States. *Water Resources Research*, *55*(4), 2916–2938. https://doi.org/10.1029/2018WR022797

Clapp, R. B., & Hornberger, G. M. (1978). Empirical equations for some soil hydraulic properties. *Water Resources Research*, *14*(4), 601–604. https://doi.org/10.1029/WR014i004p00601

Clark, M. P., & Kavetski, D. (2010). Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes. *Water Resources Research*, *46*(10), 2009WR008894. https://doi.org/10.1029/2009WR008894

Deardorff, E., Rad, A. M., Bales, J., & Flowers, T. (2022). *National Water Center Innovators Program Summer Institute Report 2022*.

Feng, D., Beck, H., Lawson, K., & Shen, C. (2023). The suitability of differentiable, physics-informed machine learning hydrologic models for ungauged regions and climate change impact assessment. *Hydrology and Earth System Sciences*, *27*(12), 2357–2373. https://doi.org/10.5194/hess-27-2357-2023

Feng, D., Fang, K., & Shen, C. (2020). Enhancing Streamflow Forecast and Extracting Insights Using Long‐Short Term Memory Networks With Data Integration at Continental Scales. *Water Resources Research*, *56*, 24. https://doi.org/ 10.1029/2019WR026793

Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy. *Water Resources Research*, *58*(10), e2022WR032404. https://doi.org/10.1029/2022WR032404

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2020). *Data for "Rainfall-Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network"* [dataset]. Zenodo. https://doi.org/10.5281/zenodo.4072701

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., & Hochreiter, S. (2021). Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrol. Earth Syst. Sci.*, *25*(4), 2045–2062. https://doi.org/10.5194/hess-25-2045-2021

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

Houska, T., Kraft, P., Chamorro-Chavez, A., & Breuer, L. (2015). SPOTting Model Parameters Using a Ready-Made Python Package. *PLOS ONE*, *10*(12), e0145180. https://doi.org/10.1371/journal.pone.0145180

Huard, D. (2023). *RavenPy Documentation*.

Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., Cunha, L. K. D., Clarke, K. C., Mazrooei, A., & Yeghiazarian, L. (2023). *Comprehensive analysis of the NOAA National Water Model: A call for heterogeneous formulations and diagnostic model selection* [Preprint]. Preprints. https://doi.org/10.22541/essoar.167415214.45806648/v1

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). *Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores* [Preprint]. Catchment hydrology/Modelling approaches. https://doi.org/10.5194/hess-2019-327

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019a). *Benchmarking a Catchment-Aware Long Short-Term MemoryNetwork (LSTM) for Large-Scale Hydrological Modeling* [Preprint]. Global hydrology/Modelling approaches. https://doi.org/10.5194/hess-2019-368

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019b). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019

La Follette, P., Ogden, F. L., & Jan, A. (2023). Layered Green & Ampt Infiltration with Redistribution. *Water Resources Research*, e2022WR033742. https://doi.org/10.1029/2022WR033742

Lahmers, T. M., Hazenberg, P., Gupta, H., Castro, C., Gochis, D., Dugger, A., Yates, D., Read, L., Karsten, L., & Wang, Y.-H. (2021). Evaluation of NOAA National Water Model Parameter Calibration in Semi-Arid Environments Prone to Channel Infiltration. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM-D-20-0198.1

Leshno, M., Lin, V. Ya., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867. https://doi.org/10/bjjdg2

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a Physically Based Hydrologic Model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

Newman, Andrew. (2014). *A large-sample watershed-scale hydrometeorological dataset for the contiguous USA* (p. approximately 2.5 GB) [Text/plain, text/tab-separated-values, png, shp]. UCAR/NCAR - GDEX. https://doi.org/10.5065/D6MW2F4D

Ogden, F. L. (n.d.). *Parameter Estimation for a Conceptual Functional Equivalen (CFE) Formulation of the National Water Model.*

Richards, L. A. (1931). CAPILLARY CONDUCTION OF LIQUIDS THROUGH POROUS MEDIUMS. *Physics*, *1*(5), 318–333. https://doi.org/10.1063/1.1745010

Richardson, L. F. (1922). *Weather Prediction by Numerical Process*. University Press.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., Li, L., Liu, X., Ren, W., Zheng, Y., Harman, C. J., Clark, M., Farthing, M., Feng, D., Kumar, P., Aboelyazeed, D., … Lawson, K. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*. https://doi.org/10.1038/s43017-023-00450-9

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration: DYNAMICALLY DIMENSIONED SEARCH ALGORITHM. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., & Shen, C. (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

van Genuchten, M. Th. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. *Soil Science Society of America Journal*, *44*(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

Western, A. W., Blöschl, G., & Grayson, R. B. (2001). Toward capturing hydrologically significant connectivity in spatial patterns. *Water Resources Research*, *37*(1), 83–97. https://doi.org/10.1029/2000WR900241

Zhang, J., Lin, P., Gao, S., & Fang, Z. (2020). Understanding the re-infiltration process to simulating streamflow in North Central Texas using the WRF-hydro modeling system. *Journal of Hydrology*, *587*, 124902. https://doi.org/10.1016/j.jhydrol.2020.124902

Zomer, R. J., Xu, J., & Trabucco, A. (2022). Version 3 of the Global Aridity Index and Potential Evapotranspiration Database. *Scientific Data*, *9*(1), 409. https://doi.org/10.1038/s41597-022-01493-1

Supplemental Information:

Table S1: Parameter configuration used for CFE calibration

Parameter			Calibrated range	
Name	Unit	Definition	Lower bound	Upper bound
bb	-	Exponent on Clapp-Hornberger (1978) function	0	21.94
smcmax	m3/m3	Maximum soil moisture content	0.20554	1
satdk	m/h	Saturated hydraulic conductivity	0	0.000726
slop	-	Slope coefficient	0	1
max_gw_storage	m	Max groundwater storage	0.01	0.25
expon	-	A primary groundwater nonlinear reservoir exponential constant	1	8
Cgw	m/timestep	A primary groundwater nonlinear reservoir constant	1.8e-6	1.8e-3
K_lf	-	Lateral flow coefficient	0	1
K_nash	-	Nash cascade discharge coefficient	0	1

Table S2: Parameters and attributes included in differentiable LGAR.

Name	Unit	Definition

α	cm^{-1}	Van Genuchten parameter representing the inverse of air-entry
n	(-)	Van Genuchten shape parameters
K_{sat}	cm/hr	Effective saturated hydraulic conductivity
θ_s	(-)	Water content at effective saturation
θ_r	(-)	Residual water content
d_p	(cm)	Maximum ponded depth
% Clay	%	Percent of soil layer composed of clay
% Silt	%	Percent of soil layer composed of silt
% Sand	%	Percent of soil layer composed of sand
pH	N/A	Soil layer pH
Organic Matter	%	Percent of organic matter within the soil layer

Table S3: Parameters and attributes included in differentiable CFE.

Name	Unit	Definition
refkdt	(-)	Runoff partitioning parameter
satdk	mh^{-1}	Saturated hydraulic conductivity
slope	(-)	Slope index
vcmx25	umol CO2 m^{-2} s^{-1}	Maximum rate of carboxylation at 25C
mfsno	(-)	Snowmelt m parmaeter
cwpvt	(-)	Empirical canopy wind parameter