StefaLand: An Efficient Geoscience Foundation Model That Improves Dynamic Land–Surface Predictions

Nicholas Kraabel[*]
Department of Civil and Environmental Engineering
Pennsylvania State University

Jiangtao Liu[*]
Department of Civil and Environmental Engineering
Pennsylvania State University

Yuchen Bian
Amazon Web Services

Daniel Kifer
Department of Computer Science & Engineering
Pennsylvania State University

Chaopeng Shen[†]
Department of Civil and Environmental Engineering
Pennsylvania State University

Stewarding natural resources, mitigating floods, droughts, wildfires, and landslides, and meeting growing demands require models that can predict climate-driven land-surface responses and human feedback with high accuracy. Traditional impact models, whether process-based, statistical, or machine learning, struggle with spatial generalization due to limited observations and concept drift. Recently proposed vision foundation models trained on satellite imagery demand massive compute and are ill-suited for dynamic land-surface prediction. We introduce StefaLand, a generative spatiotemporal earth foundation model centered on landscape interactions. StefaLand improves predictions on three tasks and four datasets: streamflow, soil moisture, and soil composition, compared to prior state-of-the-art. Results highlight its ability to generalize across diverse, data-scarce regions and support broad land-surface applications. The model builds on a masked autoencoder backbone that learns deep joint representations of landscape attributes, with a location-aware architecture fusing static and time-series inputs, attribute-based representations that drastically reduce compute, and residual fine-tuning adapters that enhance transfer. While inspired by prior methods, their alignment with geoscience and integration in one model enables robust performance on dynamic land–surface tasks. StefaLand can be pretrained and finetuned on academic compute yet outperforms state-of-the-art baselines and even fine-tuned vision foundation models. To our knowledge, this is the first geoscience land-surface foundation model that demonstrably improves dynamic land-surface interaction predictions and supports diverse downstream applications.

1 Introduction

Climate change is ushering in strong and widespread changes on the land surface, including higher frequencies of floods, droughts, wildfires and other geohazards Ebi et al. (2021); IPCC (2021). To mitigate the impact of these disasters, there are urgent needs for models that can accurately predict land surface dynamics such as streamflow, soil moisture, soil composition, landslides, snow water equivalent, groundwater levels, and vegetation carbon content. Among these, soil moisture controls the partition of rainfall into infiltration and runoff, modulates flood generation and landslides, and critically influences land-atmosphere interactions Dorigo et al. (2013). Streamflow is the flow rate of water running in the rivers, the most accessible water resource to humans, and too high or too low streamflow can cause flooding or hydrologic drought, respectively. Soil composition (sand, silt, clay fractions) governs infiltration capacity and root-zone storage, while slope–soil-vegetation interactions directly influence landslide hazards. Here, we limit our scope to the predictions of dynamical or static land surface processes that represent the impacts of climate change. Predicting these dynamic

[*] Equal contribution.

[†] Corresponding author. cshen@engr.psu.edu

variables is distinct from image-recognition tasks, as here we seek to predict what will happen in the near or distant future.

Traditionally, these tasks were undertaken by physically-based models that take atmospheric forcings (precipitation, temperature) as inputs and sequentially calculate the physical processes that eventually lead to the variables of interest Li et al. (2015). In recent years, there has been a proliferation of data-driven machine learning (ML) models Solomatine and Ostfeld (2008). These models are often set up to accept forcing (dynamic weather) and landscape characteristics (static) data as inputs, and are trained to directly predict the natural land surface variables given the weather inputs. However, up to now, most of the geoscientific ML models have been supervised ML approaches trained specifically for a narrow set of tasks. Foundation models, which are trained on broad tasks to grasp the joint data distribution, have just started to emerge in geosciences, e.g., Terramind, Prithvi and Aurora. However, they were built mainly on satellite images for landcover-identification tasks Jakubik et al. (2023; 2025); Schmude et al. (2024). There has been a notable absence of foundation models built for dynamical land surface modeling, and a lack of effort to leverage valuable temporal datasets and ground-based observations Xie et al. (2023).

A Grand challenge for geoscientific ML models is to improve their spatial generalization, because a frequent issue facing them is the sparsity and spatial imbalance of observational data. Due to the cost of installing instruments and varying policies on data sharing, data are often only available in high density in certain regions of some developed nations. For example, streamflow gauge data are available at high density in United States, Europe, Australia and Japan, but are scanty in Africa, South America, and the rest of the Asia Global Runoff Data Centre (2020). Similar distribution patterns are found for high-quality in-situ soil moisture probes and soil property measurements. Satellite data have too coarse resolution and too large uncertainty levels compared to in-situ measurements. As quantified in many studies Feng et al. (2023), a deep network trained using data from some regions can face substantial performance degradation when applied in data-scarce regions. This is partly because there are not enough sites in space to learn the true dependencies of the targets on static land surface characteristics, and partly because of systematic data discrepancies among different regions (concept drift). This means ML land surface predictions may be unusable in more than 70% of the land. This limitation is particularly acute in data-scarce regions that are often most vulnerable to climate impacts. Any method that can reliably improve spatial generalization can be a significant and rare advance to greatly democratizes information for stakeholders worldwide.

Related Work: In hydrologic and ecosystem predictions, long short-term memory (LSTM) networks Hochreiter and Schmidhuber (1997) remain the dominant architecture, in part because land surface processes often behave like Markov processes where LSTMs' gating mechanisms handle noisy continuous inputs well. Attempts to adopt transformers, so successful in natural language processing, have generally underperformed in hydrologic regression tasks Xue et al. (2023); Liu et al. (2024), with evidence of overfitting on continuous signals Zeng et al. (2022). Nonetheless, recent studies show that with task-specific modifications and careful fine-tuning, transformers can achieve competitive results and even offer advantages in autoregressive forecasting and extreme event prediction Wen et al. (2023), precisely the areas where current hydrologic models struggle most with spatial generalization.

Traditional hydrologic research on "prediction in ungauged basins" (PUB) have examined regionalization and spatial interpolation approaches including clustering or classifying catchments and transferring parameters from donor catchments in the same class Hrachowitz et al. (2013); Yang et al. (2023). Such an expert-derived design represents a crude practice of unsupervised learning that indicate the importance of understanding the joint data distribution. However, modern weakly-supervised foundation models can, in general, much better grasp the joint data distribution than expert-driven approaches. Foundation models offer a promising approach to address these spatial generalization challenges. By pretraining on large-scale datasets to learn generalizable representations, these models can potentially transfer knowledge across regions and geoscientific domains Zhang et al. (2024).

Existing geoscience foundation models such as TerraMind, Prithvi, Aurora, and AlphaEarth have largely focused on satellite imagery Hsu et al. (2024), which, while providing global coverage, may not capture the temporal dynamics and physical processes most relevant to land surface predictions. For example, TerraMind is trained on 9 million globally distributed satellite imagery samples across various modalities (optical, radar, elevation, land use) to create a generative multimodal foundation model for Earth observation Jakubik et al. (2023). It has been claimed to serve as *"any-to-any generative, multimodal foundation model for Earth observation"* Jakubik et al. (2025). However, many critical variables regarding soil and the subsurface are not directly observable from space. In addition, satellite images are noisy and a large portion of the data is redundant and irrelevant across repeated revisits. Therefore, it is unclear whether they are of high value to land-surface dynamical prediction tasks.

Our contributions: We present Spatiotemporal Earth Foundation model with Attributes for the Land Surface (StefaLand), the first land-focused geoscientific foundation model designed for dynamic land–surface prediction. StefaLand improves predictions on streamflow, soil moisture, and soil composition compared to state-of-the-art baselines. StefaLand further demonstrates strong spatial generalization across diverse landscapes and data-scarce regions. StefaLand's attribute-based rather than image-based design (with the potential to link to image-like inputs in the future) is intentionally aligned with physical processes, drastically reducing compute requirements while retaining global coverage, making it accessible to academic researchers with modest resources. Pretraining requires only about 720 GPU hours, which makes it feasible within academic budgets. The model builds on a masked autoencoder backbone, a location-aware fusion of static and time-series inputs, grouped masking to promote cross-attribute interactions, and residual fine-tuning adapters, into a coherent design guided by geoscientific knowledge. Taken together, these contributions establish StefaLand as an efficient and accessible complement to vision-based foundation models, purpose-built for hydrologic and land–surface applications.

2 Methods

Dynamic land–surface prediction requires integrating heterogeneous information: landscape attributes such as topography, soils, and geology, together with dynamic meteorological forcings like precipitation and temperature. Capturing the joint influence of these variables is essential for spatial generalization but challenging with standard architectures. StefaLand addresses this gap through an attribute-based transformer framework that unifies static and time-varying signals in a shared representation space.

2.1 Model Architecture Overview

StefaLand processes both dynamic time series and static attributes through a unified encoder-decoder architecture, and is trained as a attribute-group-based masked autoencoder (Figure D1). The model embeds multivariate inputs, applies strategic masking to force the learning of cross-domain dependencies, and reconstructs the masked portions to generate deep representations. Our location-aware transformer encoder captures spatial relationships while positional encodings maintain temporal context. The complete mathematical formulation of our model architecture is provided in Appendix A. As well as a visual representation of the model is provide in appendix D.

2.2 Pretraining Setup

Our pretraining approach uses Cross-Variable Group Masking (CVGM), which forces the model to capture interactions among groups of landscape variables (e.g., terrain, climate, soil, vegetation) rather than treating them independently. This encourages the learning of physical relationships between different environmental drivers. The objective is a reconstruction loss on the masked slice of the sequence, normalized by variable-wise standard deviations when available and with per-variable weighting via a learnable weight vector. During pretraining, the dataset is stored in shards rather than fully loaded into memory. At the start of each epoch, we randomly reindex these shards, which effectively reshuffles the available samples. This lazy-loading strategy avoids materializing the entire corpus in memory while still ensuring high sample diversity across epochs.

The pretraining dataset is a derived global attribute dataset spanning 4,229 locations over 40 years. A complete list of variables and their sources is provided in Appendix C. The accompanying code release will enable reconstruction of this dataset. Hyperparameter settings and configurations for pretraining are also listed in the appendix above.

2.3 Transformer-based Direct Prediction Models

We developed several finetuned transformer models that combine StefaLand's base structure with new units to directly predict streamflow and soil moisture. StefaLand with residual connection (resConn) is our main proposed architecture, integrating pretrained StefaLand transformers with an LSTM decoder through a residual pathway (Figure D1). This architecture enables iterative integration of pretrained transformer features throughout the decoding process, combining frozen transformer embeddings with forcings via a convolutional neural network before merging with the decoder through skip connections. This enables effective integration of pretrained features with task-specific requirements. Detailed adapter architectures are described in Appendix A.

For comparisons, "StefaLand without resConn" utilizes transformer outputs only once through a simple adapter (Figure D2), and a comparison with this model highlights the value of the resConn structure. "StefaLand Transformer" represents the base StefaLand model trained from scratch with the finetuning dataset. "StefaLand Ablation - resConn" utilizes the same structure as "StefaLand - resConn", but randomize the weights of StefaLand instead of using the pretrained weights. Comparing with these two models demonstrates the value of pretraining. As discussed below, we also implemented Terramind with resConn, which uses IBM's Terramind foundational model together with our residual connection architecture, to contrast the effectiveness of Terramind vs StefaLand pretraining. Of course, supervised LSTM serves as a widely employed and highly-competitive baseline.

2.4 Physics-Based Integration

To leverage domain knowledge, we also implemented physics-based configurations that combine our data-driven approach with process understanding. For the process-based backbone, we employed the Hydrologiska Byråns Vattenbalansavdelning (HBV) model, a conceptual hydrologic model with state variables like snow storage, soil water, and subsurface storage. In this setup, the neural network only provides parameters for the process-based backbone, which then outputs a range of interpretable variables, and they are trained together in an end-to-end fashion.

Our physics-based approaches include LSTM-HBV1.1, a baseline configuration where LSTM outputs parameters for the HBV model; StefaLand-resConn HBV1.1; and StefaLand-no resConn HBV1.1, which integrates transformer features without residual connections. These physics-based approaches maintain interpretability while leveraging improved representations from StefaLand. Detailed implementations are provided in Appendix B.

2.5 Foundation model comparisons

For completeness, we also evaluated two existing Earth-Observation-oriented foundation models, TerraMind and PrithviWxC Jakubik et al. (2025); Hsu et al. (2024), using the same fine-tuning heads and training protocols as StefaLand. TerraMind was tested on streamflow and soil moisture, and PrithviWxC on soil moisture. Inputs were harmonized to each model's modality. It is noteworthy that either Terramind or PrithviWxC requires orders of magnitude larger storage space for the input data than StefaLand, as well as computational cost for finetuning. Due to intensive data requirements for PrithviWxC on surface-level variables mostly likely to demonstrate relevance to land surface interactions, along with all 14 static variables were used. The full atmospheric variables at differing pressure levels were excluded.

3 Experiments

For all experiments described below, forcing data and static features remained consistent across all model configurations, with hyperparameters held constant within each test type except for fundamental architectural differences inherent to transformer models. Complete details of hyperparameters, forcings, and static features are provided in the Appendix C.

In addition to the neural and process-based baselines reported below, we trained plain linear regression models for every task using the same features (and appropriate standardization). These consistently underperformed all learning baselines by large margins, so we omit them from tables for readability; full metrics are provided in Appendix E.

3.1 CAMELS Streamflow Prediction

To compare the scheme's spatial generalization capability on a well benchmarked dataset, we follow the experimental setup in (Feng et al., 2021), testing prediction in ungauged basins (PUB) and ungauged regions (PUR). These correspond to randomized spatial K-fold and regional-specific K-fold regimes, reflecting real-world challenges in generalizing to nearby unmonitored basins and to data-scarce regions, respectively. We use the CAMELS dataset (Addor et al., 2017; Newman et al., 2014), restricted to the 531-basin subset with clear watershed boundaries (Newman et al., 2017). Basins were divided into 10 random groups for PUB and 7 contiguous regions for PUR, employing leave-one-out in both cases. To avoid leakage, all CAMELS-overlapping stations were removed during pretraining for PUB tests, and entire regions were excluded for PUR tests. We evaluated both direct prediction models and neural networks used to parameterize process-based models (see Methods). Direct models included StefaLand without finetuning (StefaLand-Transformer), pretrained variants with and without residual connections (resConn / noResConn), a from-scratch ablation trained only on task data, and a version using TerraMind as encoder. For process-based models, we tested LSTM HBV1.1, StefaLand-resConn HBV1.1, and StefaLand-noResConn HBV1.1.

Table 1: CAMELS Streamflow PUB and PUR Results

Model	Random holdout (ungauged basins)				Regional holdout (ungauged regions)			
	RMSE ↓	pbRMSE ↓	Corr ↑	NSE ↑	RMSE ↓	pbRMSE ↓	Corr ↑	NSE ↑
Direct Prediction Models								
LSTM - SL	1.402	1.360	0.762	0.636	1.609	1.457	0.743	0.554
StefaLand - Transformer	1.882	1.849	0.538	0.395	1.982	1.949	0.230	0.201
StefaLand - resConn	**1.111**	**1.068**	**0.869**	**0.717**	**1.344**	**1.334**	**0.801**	**0.635**
StefaLand - no resConn	1.171	1.154	0.823	0.706	1.376	1.356	0.798	0.610
Terramind - resConn	1.332	1.301	0.777	0.637	1.420	1.398	0.763	0.551
StefaLand Ablation - resConn	1.355	1.332	0.801	0.661	1.516	1.378	0.771	0.560
Physics Based Models								
LSTM - HBV1.1	1.325	1.298	0.857	0.672	1.561	1.521	0.746	0.578
StefaLand - resConn HBV1.1	**1.234**	**1.216**	**0.863**	**0.714**	**1.345**	**1.332**	**0.842**	**0.643**
StefaLand - no resConn HBV1.1	1.315	1.302	0.848	0.707	1.401	1.379	0.835	0.623
StefaLand Ablation - resConn HBV1.1	1.310	1.306	0.842	0.693	1.465	1.432	0.607	0.512

The original multi-basin LSTM studies on CAMELS defined the modern supervised learning benchmark (Kratzert et al., 2019), and subsequent large-scale comparisons confirmed that vanilla Transformers generally fail to outperform LSTMs on rainfall–runoff prediction (Liu et al., 2024, Table 1 therein). This is again reinforced in our results, where StefaLand-Transformer underperforms LSTM. The Nash-Sutcliffe model efficiency coefficient (NSE) values reported here for LSTM are highly similar to those reported in the domain literature Feng et al. (2021). Broader evaluations likewise continue to rank LSTM-family models among the best-performing approaches (Lees et al., 2021), and even Google's global flood-forecasting system adopts an encoder–decoder LSTM backbone (Nearing et al., 2024). We therefore use LSTM-SL as the supervised SOTA comparator, complemented by a process-informed LSTM–HBV1.1 variant, and additionally test TerraMind to control for the role of EO-centric pretraining.

Against these strong baselines, the foundation model pretraining clearly provides a rare boost to generalization capability compared to supervised learning models (LSTM-SL and StefaLand-Transformer) across both PUB and PUR scenarios (Table 1). StefaLand-resConn's RMSE is almost 20% lower than that of supervised LSTM in the PUB and 17% lower in the harder PUR test, while the correlation are noticebly higher. The value of pretraining is confirmed by StefaLand-Ablation-resConn, which appears to have only minor advantage compared to LSTM. The residual-connection architecture enables effective integration of transformer features with temporal dynamicsh, showing a modest benefit compared to StefaLand-no resConn, but both are clearly stronger than either LSTM or StefaLand-Transformer. Finally, Terramind-resConn has quite comparable metrics to LSTM. It appears the pretrained Terramind offers no generalization value.

The foundation model showed its versatility in supporting the parameterization of hybrid models and improving their generalization. We observe the same patterns — the pretrained StefaLand is helpful and resConn is a valuable construct. The physics-based experiments highlight that residual connections can combine pretrained representations with process-based constraints, producing more reliable spatial generalization than even the strongest LSTM configurations. We note that the hybrid model in fact showed a slightly better NSE than the direction prediction models for PUR for either LSTM-HBV1.1 or StefaLand-resConn-HBV1.1, because the physics-based equations serve as additional constraints to mitigate overfitting.

3.2 Global Streamflow

We designed a global-scale runoff prediction experiment to assess the robustness and generalization ability worldwide. We filtered global basin datasets based on data completeness and retained 3,434 basins for the global experiment. To manage computational costs efficiently, we employed random hold-out sampling combined with three-fold cross-validation. Additionally, we implemented a regionally hold-out continental scenario (RH-C), excluding all basins from North America, South America, and Europe, respectively, from the training set. The model trained from remaining continents was then evaluated on the excluded continent, simulating the realistic scenario of deploying the model to unseen geographical domains.

Table 2: Global streamflow prediction across 3,434 basins worldwide

	Random holdout (ungauged basins)			Regional holdout (ungauged continents)		
Model	RMSE ↓	μbRMSE ↓	Corr ↑	RMSE ↓	μbRMSE ↓	Corr ↑
LSTM - SL	0.870	0.864	0.798	1.253	1.202	0.672
StefaLand - Transformer	**0.749**	**0.751**	**0.843**	**1.075**	**1.048**	**0.697**
Terramind + finetuning	1.156	1.111	0.580	1.234	1.158	0.670

Cross-continental transfer testing using three-fold validation. Detailed metric calculations in Appendix E.

Results indicate that StefaLand-Transformer outperformed both LSTM-SL and the fine-tuned vision-based Transformer model (TerraMind) across all evaluation metrics. Specifically, in the RH scenario, StefaLand achieved an RMSE of 0.749, approximately 14% lower than LSTM (0.87), with Corr improving to 0.843. Notably, StefaLand's unbiased RMSE closely matched its RMSE, indicating errors were primarily random fluctuations rather than systematic bias. In contrast, TerraMind exhibited a higher RMSE of 1.156 and a low Corr of 0.580. We hypothesize that this is because the pretraining satellite image data for the vision foundation network TerraMind, did not have high relevance to hydrologic predictions.

Under the more challenging RH-C scenario, all models experienced increased errors, but StefaLand continued to demonstrate higher performance. Its RMSE was 1.075, approximately 14.1% lower than LSTM's 1.253, while also maintaining the highest Corr (0.697). Moreover, StefaLand's ubRMSE remained close to its RMSE, confirming its robust ability to correct regional-scale biases even under extreme out-of-domain conditions. Conversely, LSTM and TerraMind exhibited large gaps between ubRMSE and RMSE, highlighting their challenging in producing hydrologically-relevant features.

3.3 Global Soil Moisture

We evaluated finetuning StefaLand for soil moisture predictions following the experimental design from Liu et al. (2023a), using data from the International Soil Moisture Network (ISMN) Dorigo et al. (2011; 2013). Even though there is a globally covering satellite-based product for soil moisture, the data quality can hardly match that of in-situ moisture sensors. Thus the ability to generalize in-situ data will still be valuable. ISMN consists of 1,316 ground-based measurement stations. We randomly partitioned sites into five groups for spatial cross-validation in the random holdout test, training on four groups and testing on the fifth, rotating through all groups to obtain comprehensive spatial generalization performance. For the regional holdout test, we specifically evaluated model performance on Europe as an ungauged region, training on all other continents while excluding European sites entirely to assess cross-continental transferability, removing 129 sites for testing. We tested the following model configurations: StefaLand-Transformer (no finetuning, direct prediction), StefaLand with and without residual connections (resConn / noResConn), an ablation trained from scratch without pretraining, a baseline LSTM, and a version using IBM's TerraMind encoder with resConn.

Table 3: Soil moisture prediction across 1,316 ISMN stations

	Random location holdout (random sites)			Regional holdout (Europe)		
Model	RMSE ↓	µbRMSE ↓	Corr ↑	RMSE ↓	µbRMSE ↓	Corr ↑
LSTM - SL	0.073	0.055	0.764	0.112	**0.053**	0.510
StefaLand - Transformer	0.140	0.103	0.637	0.135	0.112	0.503
StefaLand - resConn	**0.068**	**0.054**	**0.783**	**0.090**	0.059	**0.638**
StefaLand - no resConn	0.075	0.057	0.741	0.095	0.058	0.545
Terramind - resConn	0.083	0.062	0.694	0.101	0.080	0.519
PrithviWcX - resConn	0.081	0.060	0.703	0.103	0.079	0.523
Ablation StefaLand - resConn	0.074	0.058	0.749	0.108	0.064	0.528

5-fold spatial validation and cross-continental validation on Europe (129 sites). Detailed metric calculations in Appendix E.

LSTM again serves here as the established state-of-the-art baseline. Comprehensive reviews have shown that LSTM variants consistently provide the most stable and effective performance among supervised learning methods for soil moisture prediction due to their temporal modeling strengths (Wang et al., 2024). On the ISMN-linked global dataset in particular, multitask LSTM models are widely adopted and competitive with newer architectures (Liu et al., 2023b). Given these findings, we benchmark against LSTM-SL as the SOTA/near-SOTA supervised model, and include transformer-only and vision-foundation encoders to disentangle temporal modeling from representation learning.

Against these strong baselines, the soil moisture experiments confirm the superiority of the StefaLand-resConn architecture (Table 3). StefaLand-resConn achieves the best performance, with RMSE of 0.068 and correlation of 0.783 for random holdout, and maintains strong performance even in cross-continental testing on Europe (RMSE = 0.090, Corr = 0.638). The direct StefaLand model performs poorly since transformers alone lack an effective mechanism for temporal dependencies, while the resConn architecture compensates for this limitation. Similar to the streamflow results, TerraMind and PrithviWcX both showed no performance benefit compared to LSTM despite architectural adaptations, confirming both the challenge of repurposing vision foundation models across domains and the importance of pretraining with geoscience-relevant variables. Notably, the regional holdout on Europe demonstrates StefaLand-resConn's superior spatial generalization capabilities, achieving a 25% correlation improvement over the LSTM baseline in this highly challenging extrapolation scenario.

We want to emphasize that TerraMind and PrithviWxC were not designed for the hydrologic and land–surface prediction tasks studied here. They excel in their intended domains of EO imagery and atmospheric variables, but are out-of-domain for dynamic land–surface and hydrologic modeling. We include them only to explore whether such models transfer any useful signal in our setting. Their weaker results should not be interpreted as evidence against their capability in other applications, but rather as a reflection of the mismatch between their pretraining objectives and the land–surface tasks we target. Nevertheless, we believe these benchmarks are helpful for clarifying their respective strengths, since the AI community may not be familiar with these datasets and models that have been claimed "any-to-any generative foundation model for Earth observation".

3.4 Soil Property Prediction

There are different soil datasets, each collected with different protocols and data processing techniques, resulting in significant discrepancies between datasets. In this test, we finetuned StefaLand to predict in-situ soil profile data from another dataset (ISRIC). This application can produce a seamless dataset that is consistent with a set of in-situ data, improving data availability and addressing systematic biases. In addition, it helps us understand the noise associate with each dataset. StefaLand's pretraining soils dataset is HWSD, which has some overlap but also quite extensive differences from ISRIC, which is larger and potentially noisier. We finetuned StefaLand to predict one soil texture property (e.g., clay percentage) in ISRIC while masking the corresponding complementary attributes (e.g., sand, silt) from the same profile to avoid information leakage. As a result, we are probing how easy it is to infer soil properties using other attributes such as climate, terrain and land cover. We compared StefaLand with and without pretraining against a supervised random forest baseline, which is prevalently used in the present geoscientific literature.

Table 4: In-situ soil property prediction using ISRIC WoSIS data.

Model	WoSIS Property	Corr ↑	R^2 ↑
StefaLand - Transformer	Clay	0.20	0.04
StefaLand - Finetune	Clay	**0.51**	**0.26**
Random Forest	Clay	0.05	0.00
StefaLand - Transformer	Sand	0.25	0.06
StefaLand - Finetune	Sand	**0.70**	**0.50**
Random Forest	Sand	0.26	0.07
StefaLand - Transformer	Silt	0.41	0.17
StefaLand - Finetune	Silt	0.63	0.40
Random Forest	Silt	**0.84**	**0.70**

StefaLand with finetuning achieved markedly higher predictive power, especially for sand fraction ($R^2 = 0.50$), compared to both direct training and the random forest baseline. These results highlight StefaLand's ability to reconcile disparate soil datasets and improve the utility of noisy in-situ observations. At the same time, the weaker performance on silt suggests limits of transfer: silt fractions are often residually defined (sand and clay measured directly, silt inferred as the remainder), and thus more sensitive to error propagation across datasets. This makes them inherently harder to learn, as any noise or systematic difference between HWSD and ISRIC is amplified. The fact that the random forest attains higher accuracy on silt may also indicate that simpler models can more directly exploit the compositional closure relationship, whereas StefaLand relies more on cross-variable and environmental cues, which provide less direct signal for silt.

4 Discussion

4.1 Key Findings and Contributions

As we review the literature, the methods to improve spatial generalization to data-scarce regions are rare and rarely effective Beery et al. (2018); Gacu et al. (2025). In contrast, StefaLand, combined with lightweight fine-tuning heads, achieves state-of-the-art or competitive performance across four broad problem classes: streamflow (both CAMELS and global), soil moisture, soil composition while also strengthening the parameterization of differentiable process-based models. Across tasks, the strongest gains come from architectures that fuse StefaLand embeddings with explicit temporal modeling via residual connections, indicating that pretraining on attribute-based spatiotemporal structure yields problem-relevant representations while temporal heads resolve sequence dynamics. These outcomes support the premise that foundation models can democratize prediction quality in data-scarce regions by improving out-of-domain transfer.

The streamflow experiments, spanning basin-scale CAMELS PUB/PUR and global cross-continental runoff, consistently show StefaLand's improved spatial generalization. The soil moisture results confirm robustness in both random site holdouts and continental transfer (Europe). For soil composition, StefaLand reconciles noisy WoSIS in-situ measurements with improved predictive power over random forests, highlighting its ability to harmonize disparate static datasets.

An additional contribution is computational: our attribute-based approach is markedly more efficient than pixel-wise satellite transformers. The underlying transformer has far fewer parameters (roughly 12 million) and avoids the heavy data management requirements of image-centric pretraining. TerraMind's larger configuration corresponds to about 7,680 GPU hours Jakubik et al. (2025), Aurora required roughly 14,592 GPU hours Bodnar et al. (2025), and AlphaEarth consumed on the order of 28,672 GPU hours Brown et al. (2025). These figures do not include the petabyte-scale data storage and transfer demands for satellite archives. By comparison, StefaLand's attribute-based pretraining requires only about 720 GPU hours, making high-quality spatial generalization feasible on budgets accessible to academic groups.

Finally, StefaLand is useful both operationally and as a representation layer. For hydrology, it supports point-scale prediction and regional products, for soil and geomorphic tasks it supplies frozen features that enhance conventional models. This dual role helps bridge a persistent gap between data-rich and data-poor regions by providing transferable geoscience embeddings that downstream models can reliably exploit.

4.2 Limitations

While StefaLand demonstrates strong generalization across diverse tasks, several limitations remain. The pretraining dataset is intentionally narrow and high quality, covering a carefully curated set of forcings and attributes. This focus improves signal extraction but may limit the model's usefulness for problems outside or only loosely related to those variables, such as deeper subsurface processes like groundwater dynamics. Validation is also biased toward regions with dense observational coverage, such as the United States and Europe; although regional holdout tests suggest good transfer, true performance in data-scarce regions of Africa, Asia, and South America is less certain.

4.3 Future Work

Several directions remain to further develop StefaLand. Expanding the range of targets to include variables such as evapotranspiration, snow water equivalent, and groundwater levels would broaden its applicability, while harmonized global inventories could support more comprehensive evaluation of soil and landslide predictions across differing lithologies and climates. Methodologically, advances such as sequence-aware pretraining, uncertainty-aware prediction heads, and tighter integration with process models offer promising avenues to improve calibration and interpretability while preserving efficiency. Establishing standardized train–test protocols for random and regional holdouts, released with accompanying data splits, will also be important for enabling consistent benchmarking of future geoscience foundation models.

Overall, StefaLand shows that attribute-centric pretraining combined with lightweight temporal or physics heads can deliver strong spatial generalization across geoscientific tasks while remaining computationally accessible. This points toward a practical path for high-quality predictions in regions where they are most needed but data are most limited.

5 Reproducibility Statement

The pretrained StefaLand model and all code for both pretraining and finetuning is released publicly at [https://anonymous.4open.science/r/StefaLand-9421/]. All datasets used in this work both pretraining dataset and all finetuneing benchmarks are fully public. A complete list of variables used for each task, along with their data sources, is provided in Appendix C. We also report all hyperparameters and model details in the same Appendix.

References

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

A. Aghakouchak and E. Habib. Application of a conceptual hydrologic model in teaching hydrologic processes. *International Journal of Engineering Education*, 26:963–973, 2010.

H. E. Beck, M. Pan, P. Lin, J. Seibert, A. I. J. M. van Dijk, and E. F. Wood. Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, 125:e2019JD031485, 2020. doi: 10.1029/2019JD031485.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

S. Bergström. *Development and application of a conceptual runoff model for Scandinavian catchments*. PhD thesis, Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden, 1976. URL http://urn.kb.se/resolve?urn=urn:nbn:se:smhi:diva-5738.

S. Bergström. The HBV model—its structure and applications. Technical report, Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden, 1992. URL https://www.smhi.se/en/publications/the-hbv-model-its-structure-and-applications-1.83591.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8004): 1180–1187, 2025. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025. URL https://arxiv.org/abs/2507.22291.

W. A. Dorigo, W. Wagner, R. Hohensinn, S. Hahn, C. Paulik, A. Xaver, A. Gruber, M. Drusch, S. Mecklenburg, P. van Oevelen, A. Robock, and T. Jackson. The international soil moisture network: A data hosting facility for global in situ soil moisture measurements. *Hydrology and Earth System Sciences*, 15(5):1675–1698, 2011. doi: 10.5194/hess-15-1675-2011.

W. A. Dorigo, A. Xaver, M. Vreugdenhil, A. Gruber, A. Hegyiová, A. D. Sanchis-Dufau, D. Zamojski, C. Cordes, W. Wagner, and M. Drusch. Global automated quality control of in situ soil moisture data from the international soil moisture network. *Vadose Zone Journal*, 12(3):vzj2012.0097, 2013. doi: 10.2136/vzj2012.0097.

Kristie L. Ebi, Jennifer Vanos, Jane W. Baldwin, Jesse E. Bell, David M. Hondula, Nicole A. Errett, Katie Hayes, Colleen E. Reid, Shubhayu Saha, June Spector, and Peter Berry. Extreme weather and climate change: Population health and health system implications. *Annual Review of Public Health*, 42:293–315, April 2021. doi: 10.1146/annurev-publhealth-012420-105026. PMID: 33406378; PMCID: PMC9013542.

D. Feng, H. Beck, K. Lawson, and C. Shen. 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, 2023. doi: 10.5194/hess-27-2357-2023.

Dapeng Feng, Kathryn Lawson, and Chaopeng Shen. Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, 48(12):e2021GL092999, 2021. doi: 10.1029/2021GL092999.

Jerome G. Gacu, Cris Edward F. Monjardin, Ronald Gabriel T. Mangulabnan, and Jerime Chris F. Mendez. Application of artificial intelligence in hydrological modeling for streamflow prediction in ungauged watersheds: A review. *Water*, 17(18):2722, 2025. ISSN 2073-4441. doi: 10.3390/w17182722. URL https://www.mdpi.com/2073-4441/17/18/2722.

Global Runoff Data Centre. Global runoff database, 2020. URL https://www.bafg.de/GRDC/. Accessed: 2020-04-12.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

M. Hrachowitz, H. H. G. Savenije, G. Blöschl, J. J. McDonnell, M. Sivapalan, J. W. Pomeroy, B. Arheimer, T. Blume, M. P. Clark, U. Ehret, F. Fenicia, J. E. Freer, A. Gelfan, H. V. Gupta,

D. A. Hughes, R. W. Hut, A. Montanari, S. Pande, D. Tetzlaff, P. A. Troch, S. Uhlenbrook, T. Wagener, H. C. Winsemius, R. A. Woods, E. Zehe, and C. Cudennec. A decade of predictions in ungauged basins (pub)—a review. *Hydrological Sciences Journal*, 58(6):1198–1255, 2013. doi: 10.1080/02626667.2013.803183.

C.-Y. Hsu, Wenwen Li, and S. Wang. Geospatial foundation models for image analysis: Evaluating and enhancing NASA-IBM Prithvi's domain adaptability. *International Journal of Geographical Information Science*, pages 1–30, 2024. doi: 10.1080/13658816.2024.2397441.

IPCC. Summary for policymakers. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2021.

J. Jakubik, S. Roy, C. E. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards, D. Kimura, N. Simumba, L. Chu, S. K. Mukkavilli, D. Lambhate, K. Das, R. Bangalore, D. Oliveira, M. Muszynski, et al. Foundation models for generalist geospatial artificial intelligence, 2023. URL `https://arxiv.org/abs/2310.18660`.

J. Jakubik, F. Yang, B. Blumenstiel, E. Scheurer, R. Sedona, S. Maurogiovanni, J. Bosmans, N. Dionelis, V. Marsocci, N. Kopp, R. Ramachandran, P. Fraccaro, T. Brunschwiler, G. Cavallaro, J. Bernabe-Moreno, and N. Longépé. TerraMind: Large-scale generative multimodality for Earth observation, 2025. URL `https://arxiv.org/abs/2504.11171`.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Toward learning universal, regional, and local hydrological behaviors via machine learning. *Hydrology and Earth System Sciences*, 23 (12):5089–5110, 2019. doi: 10.5194/hess-23-5089-2019.

Thomas Lees, Marcus Buechel, Bailey Anderson, Louise Slater, Steven Reece, Gemma Coxon, and Simon J. Dadson. Benchmarking data-driven rainfall–runoff models in great britain. *Hydrology and Earth System Sciences*, 25(10):5517–5534, 2021. doi: 10.5194/hess-25-5517-2021.

H.-Y. Li, L. R. Leung, A. Getirana, M. Huang, H. Wu, Y. Xu, J. Guo, and N. Voisin. Evaluating global streamflow simulations by a physically based routing model coupled with the community land model. *Journal of Hydrometeorology*, 16(2):948–971, 2015. doi: 10.1175/JHM-D-14-0079.1.

J. Liu, D. Hughes, F. Rahmani, K. Lawson, and C. Shen. Evaluating a global soil moisture dataset from a multitask model (GSM3 v1.0) with potential applications for crop threats. *Geoscientific Model Development*, 16(5):1553–1567, 2023a. doi: 10.5194/gmd-16-1553-2023.

Jiangtao Liu, David Hughes, Farzad Rahmani, Kathryn Lawson, and Chaopeng Shen. Evaluating a global soil moisture dataset from a multitask deep learning model. *Geoscientific Model Development*, 16(5):1553–1567, 2023b. doi: 10.5194/gmd-16-1553-2023.

Jiangtao Liu, Yuchen Bian, and Chaopeng Shen. Probing the limit of hydrologic predictability with the transformer network. *Journal of Hydrology*, 2024. doi: 10.1016/j.jhydrol.2024.131389.

Grey Nearing, Deborah Cohen, Vusumuzi Dube, Martin Gauch, Oren Gilon, Shaun Harrigan, Avinatan Hassidim, Daniel Klotz, Frederik Kratzert, Asher Metzger, Sella Nevo, Florian Pappenberger, Christel Prudhomme, Guy Shalev, Shlomo Shenzis, Tadele Yednkachw Tekalign, Dana Weitzner, and Yossi Matias. Global prediction of extreme floods in ungauged watersheds. *Nature*, 2024. doi: 10.1038/s41586-024-07145-1.

A. J. Newman, K. Sampson, M. P. Clark, A. Bock, R. J. Viger, and D. Blodgett. A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, 2014. URL `https://doi.org/10.5065/D6MW2F4D`. Data set.

A. J. Newman, N. Mizukami, M. P. Clark, A. W. Wood, B. Nijssen, and G. Nearing. Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, 18:2215–2225, 2017. doi: 10.1175/JHM-D-16-0284.1.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E. Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J. Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate, 2024. URL `https://arxiv.org/abs/2409.13598`.

J. Seibert and M. J. P. Vis. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16:3315–3325, 2012. doi: 10.5194/hess-16-3315-2012.

C. Shen, A. P. Appling, P. Gentine, T. Bandai, H. Gupta, A. Tartakovsky, M. Baity-Jesi, F. Fenicia, D. Kifer, L. Li, X. Liu, W. Ren, Y. Zheng, C. J. Harman, M. Clark, M. Farthing, D. Feng, P. Kumar, D. Aboelyazeed, and K. Lawson. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8):552–567, 2023. doi: 10.1038/s43017-023-00450-9.

D. P. Solomatine and A. Ostfeld. Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22, 2008. doi: 10.2166/hydro.2008.015.

Yalan Song, Kamlesh Sawadekar, Jonathan M. Frame, and Ming Pan. Physics-informed, differentiable hydrologic models for capturing unseen extreme events. ESS Open Archive, March 2025. URL `https://essopenarchive.org/doi/10.22541/essoar.172304428.82707157/v2`.

Y. Wang et al. A comprehensive study of deep learning for soil moisture prediction. *Hydrology and Earth System Sciences*, 28:917–936, 2024. doi: 10.5194/hess-28-917-2024.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey, 2023. URL `https://arxiv.org/abs/2202.07125`.

Y. Xie, Z. Wang, G. Mai, Y. Li, X. Jia, S. Gao, and S. Wang. Geo-foundation models: Reality, gaps and opportunities. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4, 2023. doi: doi/10.1145/3589132.3625616.

Weize Xue, Tianyu Li, Liang Zhou, Pengju Liu, Yu Qiao, and Lei Zhang. Make transformer great again for time series forecasting. *arXiv preprint arXiv:2305.12095*, 2023. URL `https://arxiv.org/abs/2305.12095`.

Xue Yang, Fengnian Li, Wenyan Qi, Mengyuan Zhang, Chengxi Yu, and Chong-Yu Xu. Regionalization methods for PUB: a comprehensive review of progress after the PUB decade. *Hydrology Research*, 54(7):885–900, July 2023. doi: 10.2166/nh.2023.027.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022. URL `https://arxiv.org/abs/2205.13504`.

H. Zhang, J.-J. Xu, H.-W. Cui, L. Li, Y. Yang, C.-S. Tang, and N. Boers. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–41, 2024. doi: 10.1109/MGRS.2024.3496478.

A Detailed Model Architecture

This appendix provides the complete mathematical formulation of the StefaLand model architecture.

A.1 Embedding Dynamic and Static Inputs

StefaLand independently embeds each dynamic and static variable into a latent space. Specifically, for each dynamic variable c at each time step t, a two-step nonlinear embedding is applied individually:

$$z_{t,c} = \text{GELU}(x_{t,c}W_{1,c} + b_{1,c})W_{2,c} + b_{2,c} \tag{1}$$

where $W_{1,c} \in \mathbb{R}^{1\times 64}$ and $W_{2,c} \in \mathbb{R}^{64\times 256}$ are embedding parameters. After embedding all dynamic variables individually, embeddings are stacked and summed across the variable dimension, resulting in a single embedding vector per time step:

$$z_t = \sum_{c=1}^{C} z_{t,c} \tag{2}$$

Similarly, static attributes are embedded individually:

$$z_{\text{static},i} = \text{GELU}(s_i W_{1,i} + b_{1,i}) W_{2,i} + b_{2,i} \tag{3}$$

where separate embedding layers are used for static features. These individual static embeddings are then concatenated with dynamic embeddings along the temporal dimension, resulting in a unified embedding tensor:

$$Z = [z_1; z_2; \ldots; z_T; z_{\text{static}}] \tag{4}$$

This static embedding acts as a global learnable token, allowing the model to incorporate basin-specific context into temporal dynamics at any depth of the Transformer layers.

A.2 Transformer Encoder

The embeddings enriched by positional encoding are processed through an N-layer Transformer encoder, where each Transformer block successively applies Multi-Head Self-Attention (MHA) with h attention heads, followed by a residual connection and Layer Normalization. Subsequently, a position-wise Feedforward Network (FFN) is applied, also followed by another residual connection and Layer Normalization:

$$\begin{align} A^{(\ell)} &= \text{MHA}(H^{(\ell-1)}) \tag{5}\\ \tilde{H}^{(\ell)} &= \text{LayerNorm}(H^{(\ell-1)} + A^{(\ell)}) \tag{6}\\ F^{(\ell)} &= \text{FFN}(\tilde{H}^{(\ell)}) \tag{7}\\ H^{(\ell)} &= \text{LayerNorm}(\tilde{H}^{(\ell)} + F^{(\ell)}) \tag{8} \end{align}$$

A.3 Reconstruction of Original Inputs

The final hidden states from the Transformer encoder, $H^{(N)}$, are linearly projected and passed through a single-layer bidirectional LSTM to capture the local temporal dependencies and continuity:

$$U = \text{LSTM}(H^{(N)} W_{\text{enc-proj}} + b_{\text{enc-proj}}) \tag{9}$$

The outputs U are then separated into dynamic and static components, U_t and U_{static}, corresponding to the temporal sequence and static attributes:

$$U_t, U_{\text{static}} = U_{1:T}, U_{T+1} \tag{10}$$

Finally, both dynamic and static representations are individually projected back to their original dimensions through separate embedding layers, reconstructing the masked portions of the inputs. Dynamic variables are restored via:

$$\hat{x}_t = \text{DynamicDecEmbedding}(U_t) \tag{11}$$

while static attributes are restored by:

$$\hat{s} = \text{StaticDecEmbedding}(U_{\text{static}}) \tag{12}$$

The projections leverage the learned latent representations to reconstruct the original hydrologic inputs.

B Physics-Based Differentiable Modeling

To leverage domain knowledge and physical constraints inherent in hydrological systems, we implemented physics-based models that explicitly represent hydrological processes through mathematical formulations. These differentiable versions can be trained end-to-end within neural network frameworks, combining process understanding with machine learning flexibility Shen et al. (2023).

For the process-based backbone, we employed the Hydrologiska Byråns Vattenbalansavdelning (HBV) model Aghakouchak and Habib (2010); Beck et al. (2020); Bergström (1976; 1992); Seibert and Vis (2012), a relatively simple bucket-type conceptual hydrologic model. HBV has state variables like snow storage, soil water, and subsurface storage, and can simulate flux variables like evapotranspiration (ET), recharge, surface runoff, shallow subsurface flow, and groundwater flow. The parameters of HBV are learned from basin characteristics by a deep learning network.

We used an updated modern version, HBV1.1 Song et al. (2025), which includes modifications such as increased parallel storage components to represent heterogeneity within basins and dynamic parameterization capabilities.

For physics-based configurations, we tested: (1) a baseline LSTM-HBV1.1 configuration as a standard reference, (2) StefaLand HBV1.1 with resConn, which combines the physics-based approach with our residual connection architecture, and (3) StefaLand HBV1.1 without resConn. These physics-based approaches incorporate hydrological process understanding while maintaining the ability to learn from data.

Table C1: StefaLand Pretraining Variables and Sources

Variable Type	Variable Name	Source
Time Series Forcings	Precipitation, Short-wave solar radiation downwards, Relative humidity, Maximum temperature, Minimum temperature, Potential evapotranspiration	Multi-Source Weather (MSWX) and Multi-Source Weighted-Ensemble Precipitation (MSWEP)
Static Attributes	Forest cover fraction, grassland cover fraction	Climate Change Initiative (CCI) land cover dataset
	Normalized Difference Vegetation Index (NDVI)	Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13A3)
	Sand, silt, clay fractions	Harmonized World Soil Database (HWSD)
	Elevation, slope, aspect	Global Multi-resolution Terrain Elevation Data (GMTED)
	Soil depth	Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers
	Carbonate sedimentary rock fraction	Global Lithological Map (GLiM)
	Rock porosity, permeability	GLobal HYdrogeology MaPS (GLHYMPS)
	Population density	Gridded Population of the World (GPW) v4 dataset
	GDP per capita; population density	Gross Domestic Product and Human Development Index over 1990-2015
	Forest intact fraction	Intact Forest Landscapes Data
Outputs	None (self-supervised pretraining)	—

C EXPERIMENTAL DETAILS

Table C2: CAMELS Streamflow HBV Model Hyperparameters

Parameter	Value
General Settings	
Random seed	111111
Data sampler	finetune_sampler
Training Configuration	
Time period	1989/10/01–2008/09/30
Optimizer	Adadelta
Batch size	64
Epochs	25
Neural Model Configuration	
Sequence length	365
Hidden size	512
Dropout	0.2
Encoder layers	4
Decoder layers	2
Feed-forward dimension	512
Physical Model (HBV-1.1)	
Model type	HBV_1_1p
Number of runs (nmul)	16
Warm-up period	365 days
Warm-up states	True
Dynamic dropout	0.0
Use routing	True
Dynamic parameters	parBETA, parK0, parBETAET
Near-zero threshold	1e-05
Loss Function	
Type	RmseLoss

Table C3: CAMELS Streamflow Variables and Sources

Variable Type	Variable Name	Source
Time Series Forcings	Precipitation, Temperature, Potential evapotranspiration, Solar radiation, Vapor pressure	Catchment Attributes and Meteorology for Large-sample Studies (CAMELS)
Static Attributes	Elevation, slope, catchment area, forest cover, LAI, GVF, soil depth, porosity, conductivity, sand, silt, clay fractions, carbonate fraction, permeability, aridity, snow fraction, precipitation extremes	CAMELS
Outputs	Streamflow	CAMELS gauge records

Table C4: Soil Moisture Model Configuration

Parameter	Value
General Settings	
Mode	train_test
Random seed	111111
Data loader	onlylstm_loader
Data sampler	finetuning_noHBV
Training Configuration	
Time period	2015/04/01–2020/12/31
Target	soil_moisture
Optimizer	Adadelta
Batch size	128
Epochs	50
Save frequency	Every 25 epochs
Neural Network Configuration	
Hidden size	128
Dropout	0.3
Learning rate	1.2
Encoder layers	16
Decoder layers	12
Feed-forward dimension	512
Rho	365
Loss Function	
Type	RmseLoss

Table C5: Soil Moisture Variables and Sources

Variable Type	Variable Name	Source
Time Series Forcings	Albedo (BSA, WSA)	Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A3 Version 6
	LST (Day, Night)	MODIS Land Surface Temperature/Emissivity Daily (MYD11A1) Version 6.1
	Precipitation	Global Precipitation Measurement (GPM), Multi-Source WeightedEnsemble Precipitation (MSWEP)
	Forecast albedo, LAI (high/low vegetation), soil temperature (layer 1), surface pressure, solar radiation, 2m temperature, evaporation, precipitation, U/V wind (10m)	ERA5-Land Hourly - ECMWF Climate Reanalysis
Static Attributes	elevation, slope, aspect, roughness, curvature	Global 1,5,10,100-km Topography database
	Sand, clay, silt, bulk density	Harmonized world soil database (HWSD)
	Land cover; urban; open water; snow/ice	ESA CCI land cover 2018
	NDVI	Vegetation Indices Monthly L3 Global 0.05Deg CMG
Outputs	Soil moisture	ESA CCI SM / in-situ merged

Table C6: Global Streamflow Variables and Sources

Variable Type	Variable Name	Source
Time Series Forcings	Precipitation, Temperature, Potential evapotranspiration, Radiation, Humidity	MSWX and MSWEP
Static Attributes	Forest cover fraction, grassland cover fraction	Climate Change Initiative (CCI) land cover dataset
	Normalized Difference Vegetation Index (NDVI)	Terra Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices (MOD13A3)
	Sand, silt, clay fractions	Harmonized World Soil Database (HWSD)
	Elevation, slope, aspect	Global Multi-resolution Terrain Elevation Data (GMTED)
	Soil depth	Global 1-km Gridded Thickness of Soil, Regolith, and Sedimentary Deposit Layers
	Carbonate sedimentary rock fraction	Global Lithological Map (GLiM)
	Rock porosity, permeability	GLobal HYdrogeology MaPS (GLHYMPS)
	Population density	Gridded Population of the World (GPW) v4 dataset
	GDP per capita; population density	Gross Domestic Product and Human Development Index over 1990-2015
	Forest intact fraction	Intact Forest Landscapes Data
Outputs	Streamflow	Global gauge networks

Table C7: Soil Composition (ISRIC) Variables and Sources

Variable Type	Variable Name	Source
Time Series Forcings	Same as Table 7	—
Static Attributes	Same as Table 7	
Outputs	Soil property (clay; sand; silt)	The World Soil Information Service (WoSIS)

Table C8: Computation Resources for StefaLand and Comparison Models

Model	**Seconds/Epoch**	**#GPUs**	**GPU Type**	**Memory**
StefaLand (Pretraining)	16,000	6	NVIDIA V100	240 GB
StefaLand with resConn	30	2	NVIDIA V100	80 GB
StefaLand without resConn	26	2	NVIDIA V100	80 GB
LSTM Baseline	12	2	NVIDIA V100	80 GB
LSTM-HBV1.1	280	2	NVIDIA V100	80 GB
StefaLand-resConn HBV1.1	320	2	NVIDIA V100	80 GB
StefaLand-no resConn HBV1.1	300	2	NVIDIA V100	80 GB

Note: All values except pretraining are for the CAMELS benchmark experiment. The relative differences in computational requirements are consistent across other experiments.

D Model Architectures

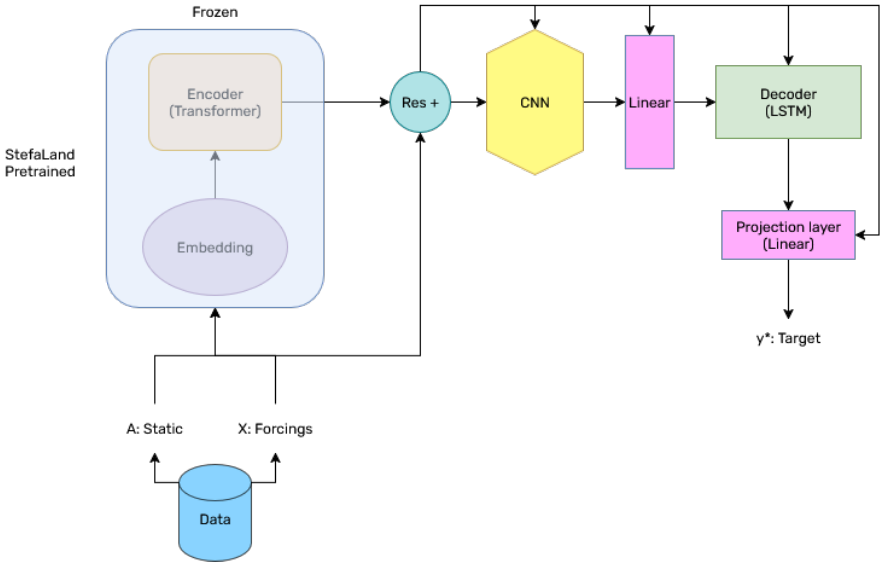

Figure D1: StefaLand model architecture with residual connections. The pretrained StefaLand encoder (frozen during fine-tuning) processes static attributes and generates embeddings that are combined with meteorological forcings through residual connections. This architecture enables iterative integration of transformer features with input data through the CNN module, allowing the model to effectively leverage pretrained representations while adapting to task-specific requirements.

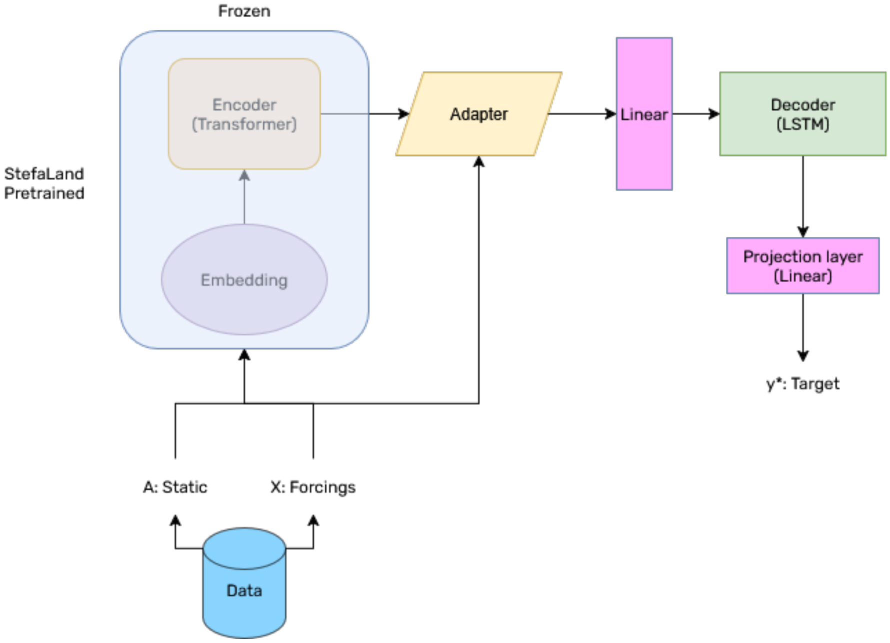

Figure D2: StefaLand model architecture without residual connections. In this configuration, the frozen transformer embeddings are used only once through a standard adapter that combines them with forcings and static features. This represents a more conventional fine-tuning approach where transformer features are not integrated iteratively throughout the decoding process.

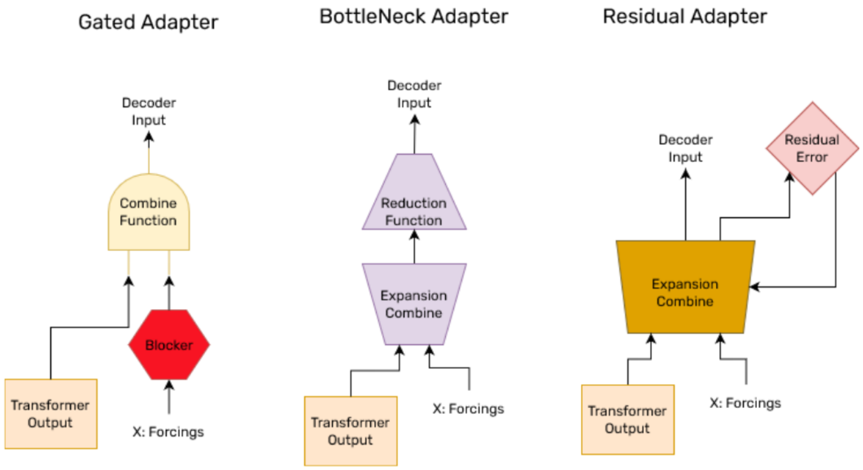

Figure D3: Different adapter architectures tested in our experiments. Left: Gated Adapter with a selector mechanism that controls information flow. Center: BottleNeck Adapter with compression and expansion phases. Right: Residual Adapter that adds transformer features through a skip connection.

E Metric Calculations

This appendix details the calculation of the evaluation metrics used in our experiments. All metrics presented in the main paper tables are the median values across test basins or stations, as computed using the following formulations.

E.1 Primary Evaluation Metrics

E.1.1 Root Mean Square Error (RMSE)

RMSE measures the average magnitude of prediction errors. Lower values indicate better performance.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{pred},i} - y_{\text{target},i})^2} \tag{13}$$

E.1.2 Unbiased Root Mean Square Error (µbRMSE)

µbRMSE removes the bias component from the error calculation, focusing on the error's random component. It is calculated by first computing anomalies from the mean for both predictions and targets.

$$y'_{\text{pred},i} = y_{\text{pred},i} - \overline{y}_{\text{pred}} \tag{14}$$

$$y'_{\text{target},i} = y_{\text{target},i} - \overline{y}_{\text{target}} \tag{15}$$

$$\text{µbRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'_{\text{pred},i} - y'_{\text{target},i})^2} \tag{16}$$

E.1.3 Correlation (Corr)

Correlation quantifies the linear relationship between predictions and targets. Values range from -1 to 1, with 1 indicating perfect positive correlation.

$$\text{Corr} = \frac{\sum_{i=1}^{n}(y_{\text{pred},i} - \overline{y}_{\text{pred}})(y_{\text{target},i} - \overline{y}_{\text{target}})}{\sqrt{\sum_{i=1}^{n}(y_{\text{pred},i} - \overline{y}_{\text{pred}})^2 \sum_{i=1}^{n}(y_{\text{target},i} - \overline{y}_{\text{target}})^2}} \tag{17}$$

This is calculated using Pearson's correlation coefficient between predicted and observed values.

E.2 Secondary Metrics

The following metrics are used in our comprehensive evaluation but may not appear directly in the main tables.

E.2.1 Nash-Sutcliffe Efficiency (NSE) / R^2

NSE evaluates the predictive skill relative to using the mean of observations as a predictor. Values range from $-\infty$ to 1, with 1 indicating perfect prediction.

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_{\text{target},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{n}(y_{\text{target},i} - \overline{y}_{\text{target}})^2} \tag{18}$$

E.2.2 Mean Absolute Error (MAE)

MAE measures the average absolute difference between predictions and targets.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_{\text{pred},i} - y_{\text{target},i}| \tag{19}$$

E.2.3 Flow Duration Curve RMSE (RMSE_FDC)

RMSE_FDC evaluates errors in the statistical distribution of flows rather than in their timing.

$$\text{RMSE\_FDC} = \sqrt{\frac{1}{100} \sum_{j=1}^{100} (FDC_{\text{pred},j} - FDC_{\text{target},j})^2} \tag{20}$$

where FDC_j represents the j-th percentile of the sorted flow values.

E.2.4 Flow Biases

Several flow-specific biases were computed to evaluate performance across different flow regimes:

- FLV (Low Flow Volume Bias): Percent bias in the lowest 30% of flows
- FHV (High Flow Volume Bias): Percent bias in the highest 2% of flows
- $PBIAS$ (Percent Bias): Overall percent bias across all flows

The general form for these biases is:

$$\text{PBIAS}_{\text{regime}} = \frac{\sum(y_{\text{pred,regime}} - y_{\text{target,regime}})}{\sum y_{\text{target,regime}}} \times 100\% \tag{21}$$

E.2.5 Kling-Gupta Efficiency (KGE)

KGE combines correlation, bias, and variability components:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\text{pred}}}{\sigma_{\text{target}}} - 1\right)^2 + \left(\frac{\mu_{\text{pred}}}{\mu_{\text{target}}} - 1\right)^2} \tag{22}$$

where r is the correlation coefficient, σ represents standard deviation, and μ represents the mean.

E.3 Metric Aggregation

For each evaluation scenario (Random Holdout and Regional Holdout), metrics were calculated for each individual basin or station and then aggregated using median values to provide a robust measure of central tendency less sensitive to outliers. All metrics shown in tables throughout the paper represent these median values across the test set.

E.4 Implementation Details

All metrics were implemented in Python using NumPy for numerical computations and SciPy's statistical functions for correlation coefficients. Special care was taken to handle missing values (NaNs) appropriately in all calculations. For time series with missing values, only timestamps where both predicted and target values were available were used in metric calculations.

F Additional Model results

Table F1: Additional experiments with linear regression baselines.

Experiment	Random holdout			Regional holdout		
	RMSE ↓	µbRMSE ↓	Corr ↑	RMSE ↓	µbRMSE ↓	Corr ↑
Camels Streamflow Linear Regression	2.190	2.180	0.500	2.260	2.250	0.500
Global Streamflow Linear Regression	1.823	1.746	0.252	1.816	1.721	0.248
Soil Moisture Linear Regression	0.120	0.101	0.188	0.121	0.103	0.187