EPiC Series in Technology

Volume 3, 2026, Pages 16–26

Proceedings of AI for Accelerated Research Symposium

FREE: The Foundational Semantic Recognition for Modeling Environmental Ecosystems

Shiyuan Luo[1], Juntong Ni[2], Shengyu Chen[1], Runlong Yu[3], Yiqun Xie[4], Licheng Liu[5], Zhenong Jin[5], Huaxiu Yao[6], and Xiaowei Jia[1]

[1] University of Pittsburgh, Pittsburgh, PA, USA
{shl298, shc160, xiaowei}@pitt.edu
[2] Emory University, Atlanta, GA, USA
juntongni02@gmail.com
[3] University of Alabama, Tuscaloosa, AL, USA
ryu5@ua.edu
[4] University of Maryland, College Park, MD, USA
xie@umd.edu
[5] University of Minnesota, Minneapolis, MN, USA
{lichengl, jinzn}@umn.edu
[6] University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
huaxiu@cs.unc.edu

Abstract

Modeling environmental ecosystems is critical for the sustainability of our planet, but is extremely challenging due to the complex underlying processes driven by interactions amongst a large number of physical variables. As many variables are difficult to measure at large scales, existing works often utilize a combination of observable features and locally available measurements or modeled values as input to build models for a specific study region and time period. This raises a fundamental question in advancing the modeling of environmental ecosystems: *how to build a general framework for modeling the complex relationships among diverse environmental variables over space and time?* In this paper, we introduce a framework, FREE, that enables the use of varying features and available information to train a universal model. The core idea is to map available environmental data into a text space and then convert the traditional predictive modeling task in environmental science to a semantic recognition problem. Our evaluation on two societally important real-world applications, stream water temperature prediction and crop yield prediction, demonstrates the superiority of FREE over multiple baselines, even in data-sparse scenarios.

1 Introduction

Understanding the dynamics of environmental ecosystems is critical for the sustainable management of natural resources and mitigating natural disasters such as algal blooms and floods, especially given the compelling need for food and water supply from a growing world population and a more unpredictable climate. Modeling environmental ecosystems is challenging as these

systems are shaped by the complex interactions of a large number of physical variables, such as weather, soil, water, and plants. Hence, it often requires the combination of data from the network of weather stations, remote sensing, and field measurements to jointly model complex system dynamics. However, many physical variables are often only sparsely available at certain locations or during certain time periods due to the substantial cost required for the data collection.

We focus on two important applications: predicting stream water temperature and annual crop yield under changing weather conditions. Success in both prediction tasks can support a range of applications in optimizing resource allocation and management strategies. For example, water reservoir operators in the Delaware River Basin (DRB) need to maintain water temperature within a desired range to ensure safe drinking water for over 15 million people while also preserving sufficiently cool water for aquatic life in downstream areas [18]; the changing climate is making it making it increasingly difficult to sustain high crop yields in the U.S. Corn Belt, threatening food supplies and farmer livelihoods. While physics-based models have been developed to simulate underlying physical processes for these ecosystems [20], the majority of them are necessarily approximations of reality due to incomplete knowledge or excessive complexity in modeling certain processes [6]. Machine learning (ML) models offer an alternative, given their computational efficiency and ability to automatically extract complex data patterns [7, 12, 8].

However, traditional ML approaches face several major challenges in fully leveraging available data for modeling ecosystems: (1) treating physical variables as independent numerical features without explicitly capturing their interdependent physical and ecological relationships. This is further exacerbated by the scarcity of observation data in many real-world ecosystems, which limits the ability of ML models to automatically extract generalizable relationships between features. (2) inability to harness inconsistent feature inputs. Current studies on local environmental ecosystems often enhance the prediction of their target study region by leveraging multiple types of input features related to the target variable, and the selected features often vary between different studies. Besides meteorological data (e.g., solar radiation) that are commonly used, prior works also explored including other measurements on the environment system (e.g., soil properties [3], land use and geometric structures [4]), and other physical variables simulated by physics-based models [9]. These features may not always be available for data samples collected from different locations and time periods, which poses a challenge in training a global model that can utilize different input features. (3) lack of a general pipeline that can dynamically incorporate auxiliary observations into the ML model. Current works on assimilating auxiliary observations rely on task-specific learning mechanisms or model structures, e.g., Kalman filtering for incorporating new observations [21], and graph convolution and invertible network layers for assimilating observations from neighboring samples [2, 1]. Yet these approaches can be computationally expensive to implement when we have a large state space and/or non-linear system dynamics. Recent LLM advances show promise for addressing these limitations [10]. Originally developed for language tasks, LLMs now handle tabular data by converting it to text [17, 19], enabling flexible feature handling and robustness to missing values.

In this paper, we propose a novel method, **F**oundational semantic **R**ecognition for modeling **E**nvironmental **E**cosystems (FREE) to address these limitations. The key idea of FREE is to translate the heterogeneous input data into natural language descriptions using large language models and then estimate the target variable through semantic recognition in the text space. This addresses the challenges in utilizing different data sources by only manipulating the text space while maintaining the same predictive modeling component (i.e., semantic recognition). In particular, the translation process uses the available features for each data point, allowing the set

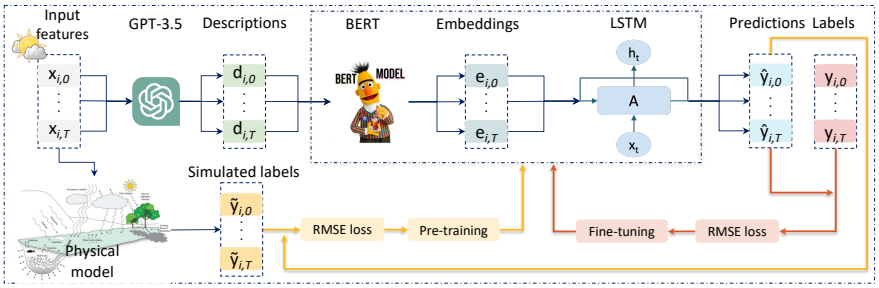

Figure 1: The framework of FREE: Input features are first transformed into natural language descriptions by LLMs. These descriptions are then processed by a LM to generate embeddings, which are fed to an LSTM layer for making predictions. Simulated labels generated by a physics-based model are used to pre-train the LM and LSTM layers, followed by fine-tuning with true observations of the target variable for enhanced predictions.

of features to vary across data points. Moreover, the translation process can easily incorporate auxiliary observations (e.g., newly collected observations from the previous day) into the textual description with a properly designed prompt. After obtaining the textual description, we build the semantic recognition component by leveraging a separate language model to embed the text data and incorporating additional network layers (e.g., long-short term memory (LSTM)) to capture temporal dependencies. The use of the language model on textual descriptions enables a better understanding of the nature and semantics of input features. To further enhance the embedding performance of the language model (LM) on environmental descriptions, we pre-train the semantic recognition component using abundant simulated samples generated by physics-based models. This pre-training process also helps the model learn the general physical relationships encoded in the physics-based models and mitigate the challenge posed by the sparse observations.

We evaluate FREE on two real-world tasks: daily stream water temperature prediction in the DRB, and annual county-wise corn yield prediction in Illinois and Iowa. Both of them cover a diverse set of locations and long time periods. FREE shows its superior predictive performance over multiple baselines, especially with sparse observations, and it can effectively handle different input features and incorporate auxiliary observations. The pre-training process also helps improve the prediction performance of the model when adapted to different locations. This work also opens new opportunities to leverage LLMs as a universal tool for general environmental modeling tasks, which are currently often addressed in a fragmented and localized manner due to variations in data collection and processing pipelines.

2 Method

In this section, we define the problem and introduce our proposed framework. The objective is to predict the target variable y (e.g., water temperature) at multiple locations $i \in \{1, ..., N\}$ and over a period of time. For clarity, we denote by $\mathbf{x}_{i,t}$ the input features at the location i on date t. The input features contain the set of meteorological variables (e.g., solar radiation, rainfall) that are commonly used as drivers for physics-based models. Besides, we include some other physical variables estimated by physics-based models (e.g., cloud cover fraction, groundwater properties) or obtained from other data sources (e.g., soil properties). Notably, these features

may be absent for certain locations or time steps. We represent the observations of the target variable at each location i over multiple time steps as $y_i = \{y_{i,1}, y_{i,2}, ..., y_{i,T}\}$. The observations can be only sparsely available for certain time steps and locations.

FREE converts the traditional predictive modeling task in environmental science to the semantic recognition problem in a text space created by LLMs (e.g., GPT). A clear benefit of this framework is its ability to harness inputs of different feature sets and incorporate auxiliary information. We also introduce additional model components to embed the text data and capture the temporal dependencies. To mitigate the need for large training samples for different ecosystems, we pre-train the model using abundant simulated data generated by physics-based models. In the following, we will describe these components in detail.

2.1 Overall architecture

The FREE framework (shown in Fig. 1) consists of two major components in its architecture.

Input data conversion: To address inconsistencies in the feature set and incorporate auxiliary observations, we propose to transform the original data sample into a corresponding natural language description. This approach facilitates the handling of diverse and potentially incomplete feature sets for different data points, enabling a uniform textual representation of data across varying input scenarios. Specifically, we leverage an existing LLM to transform each data point $\mathbf{x}_{i,t}$ into clear, natural language descriptions $d_{i,t}$. To effectively communicate with the LLM, we use a linearization technique [17] to construct prompts that consist of a context-setting prefix, the linearized data input, and a directive suffix. The prefix (p) provides the model with a background of the dataset, while the suffix (s) instructs the model on how to format its output. The complete prompt is formulated as:

$$d_{i,t} = \text{LLM}\left(p, \text{linearize}(\mathbf{x}_{i,t}), s\right). \quad (1)$$

As depicted in Fig. 2, we format the input features $\mathbf{x}_{i,t}$ of K dimensions/variables into a sequence pairing column names $c^k_{i,t}$ with the corresponding feature values $\mathbf{x}^k_{i,t}$ (e.g., { [rainfall: 0], [solar radiation: 151.14] }). This process can be expressed as follows:

$$\text{linearize}(\mathbf{x}_{i,t}) = \{[c^k_{i,t} : \mathbf{x}^k_{i,t}]\}^K_{k=1}. \quad (2)$$

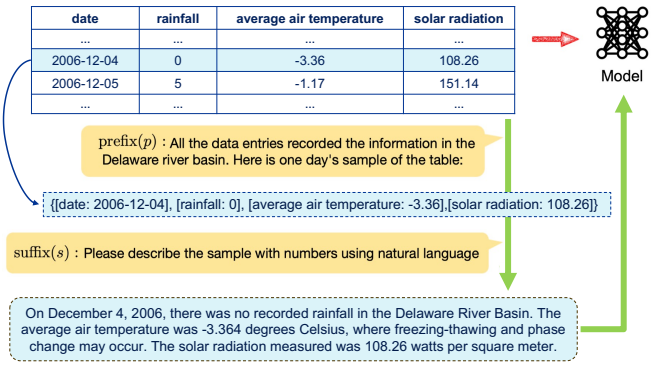

Figure 2: Green arrows indicate FREE handling inputs of diverse feature sets that use linearized data. The red arrow suggests that traditional ML models may require separate preprocessing methods to address data irregularities.

Given the prompt, LLM can generate feature summaries that are both descriptive and succinct. , such as: "*On December 4, 2006, there was no recorded rainfall in the Delaware River Basin. The average air temperature was -3.36 degrees Celsius, where freezing-thawing and phase change may occur. The solar radiation measured was 108.26 watts per square meter.*". The obtained textual descriptions, alongside the observations in the original dataset, form paired data samples as $\{d_{i,t}, y_{i,t}\}$, which are then used for tuning the semantic recognition component (to be discussed later).

In summary, the proposed input translation process helps generate an understandable and flexible representation of input features. More importantly, the use of LLM enables supplementing

the raw feature values with semantically meaningful text descriptions and interpreting underlying physical phenomena, which facilitates subsequent modeling components to capture the complex nature and the interactions of input features.

Semantic recognition: Once obtaining the textual descriptions $d_{i,t}$, we utilize a separate LM f to process the obtained text data and embed them as

$$\mathbf{e}_{i,t} = f(d_{i,t}). \tag{3}$$

In this work, we use the DistilBERT model [13] in this step. It is vital to note that the embeddings $\mathbf{e}_{i,t}$ are generated independently across different data points. To effectively capture the intricate data temporal dependencies, we introduce additional network layers that combine multiple data embeddings. In particular, this work uses LSTM layers to capture common temporal data dependencies in the environmental ecosystem, e.g., seasonal changes, and the effect of heavy rainfall on soil water in the next few days.

2.2 Handling diverse inputs and auxiliary data

One major goal of modeling environmental ecosystems is long-term prediction in large regions. FREE provides opportunities to enhance ML models towards this goal from two aspects. First, predictive models often have larger predictive errors as time progresses in the test phase due to the accumulated bias in the model state. One could explore leveraging newly collected observations during the testing phase (e.g., the observations collected at previous time $t-1$) to mitigate the current model bias and improve the future predictions (e.g., the prediction after t). Second, when training the model across space, some areas may offer more features, such as certain variables related to local soil properties while other regions may only provide weather-related features. Such discrepancies in feature availability pose challenges for building a universal model across space. Under the FREE framework, we can leverage additional observations and handle different features through simple modification of the prompt while keeping the predictive model (i.e., the semantic recognition component) unchanged.

Incorporating auxiliary observations: In this study, we consider incorporating two types of auxiliary observations to enhance the prediction of the data sample $\mathbf{x}_{i,t}$: the newly collected observation from the previous time $t-1$ and at the same location i, and from the neighboring locations of the target location i. We incorporate the auxiliary data by updating the linearized data input to include it. As observations are collected from different dates with the existing features in $\mathbf{x}_{i,t}$, we need to explicitly include the exact date information for the auxiliary observations and original features. If we consider the auxiliary observations being collected at the date $t-1$ from both the current location i and its neighbor $j \in \mathcal{N}(i)$, the updated linearization process can be expressed as:

$$\begin{aligned} &\text{linearize}([\mathbf{x}_{i,t}, y_{i,t-1}, y_{j,t-1}]_{j\in\mathcal{N}(i)}) \\ =&[\text{date} : t-1] \cup [c_i^y : y_{i,t-1}] \cup [c_j^y : y_{j,t-1}] \cup [\text{date} : t] \cup \{[c_{i,t}^k : \mathbf{x}_{i,t}^k]\}_{k=1}^K, \end{aligned} \tag{4}$$

where c_i^y denotes the column name for the observed labels, c_j^y denotes the column name for the observed labels from the neighboring location j, and $\cup$ indicates concatenation operation across the sequences. LLM then follows the instructions by p and s to generate the description using these data. It is noteworthy that the auxiliary observations may not always be available. When creating the linearized data, we include the columns of the auxiliary observations only if they are available, and skip them otherwise.

Handling different input features: FREE handles different input features by linearizing only available features, skipping missing ones. In particular, if a feature k is not available for the

data sample $\mathbf{x}_{i,t}$, then the pair $[c^k_{i,t} : \mathbf{x}^k_{i,t}]$ will be skipped and not included in the input to the LLM. This enables the use of a combination of heterogeneous data samples with different feature sets in both training and testing processes. This also allows the subsequent semantic recognition to proceed seamlessly on the generated text without the need for manual adjustments to account for data irregularities.

2.3 Pre-training using physical simulations

The LM in the semantic recognition phase (Eq. 3) are not inherently trained on data specific to the target environmental ecosystems. As shown in Fig. 3 (b), the input features samples from different seasons tend to be mixed in the latent embedding space of the original LM, when applied directly, may fall short of capturing semantics from the text generated for our target task. Tuning the LM towards the target domain requires sufficient observed samples, which are often not available in real-world ecosystems. Therefore, we use simulated data $\tilde{y}_{i,t}$ generated by physics-based model to pre-train the semantic recognition component. Fig. 3 (a) shows that samples of different seasons can now be distinguished, confirming that the LM model tuned with simulated data can better capture semantics in the textual descriptions $d_{i,t}$. This simulation-based pre-training also facilitates the training of the semantic recognition component to emulate general physical processes encoded in the physics-based model, enhancing the model's generalizability as many physical processes generally hold across space and time. Upon completing this pre-training, it requires only a few epochs of fine-tuning using true observations before reaching a quality model.

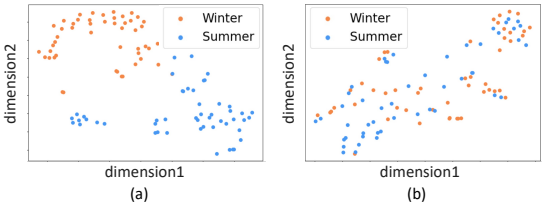

Figure 3: t-SNE of embeddings for data points randomly sampled from summer and winter.

3 Experiments

In this section, we present our datasets and provide a comprehensive assessment of the proposed methods. Our primary emphasis lies on predictive performance using sparse data, underscoring the efficacy of FREE framework and the simulation-based pre-training. Besides, we delve into additional experiments that validate our model's capability to harness diverse input features and auxiliary observations. We compare FREE with multiple popular baselines developed for stream modeling, including LSTM model (considered to be the state of the art in many environmental modeling problems [14]), Transformer [16], RGRN [9],Gr-CNN [15], HydroNets [11]. Besides, we include a baseline that assimilates new observations for updating the model, HRGN-DA [2]. We compare the proposed method with the LSTM and two other baselines that have both proven effective in crop yield prediction, i.e., PG-AN [7], and PG-GNN [5].

3.1 Overall predictive performance

Tables 1 and 2 show the performance of various methods for predicting stream water temperature and annual crop yield, respectively. The notation FREE represents a model pre-trained on synthetic data and then fine-tuned on observational data. Table 1 shows that FREE outperforms all baselines when data are abundant, showing its effectiveness

Table 1: Prediction RMSE for stream temperature using 1%, 2%, 4%, and 100% randomly selected training labels. The best results are **bold**, the second best results are underlined.

Dataset	Method	100%	4%	2%	1%
SS	LSTM	1.90	1.95	2.14	2.23
	Transformer	1.79	2.05	2.20	2.27
	FREE	**1.70**	**1.71**	**1.90**	**2.09**
CRW	LSTM	1.90	2.33	2.80	2.87
	Transformer	2.14	2.20	2.19	2.30
	RGRN	1.78	2.31	2.61	2.80
	Gr-CNN	1.80	2.38	2.69	3.56
	HydroNets	1.87	2.51	2.77	3.75
	FREE	**1.61**	**1.68**	**1.62**	**1.65**

Table 2: Prediction RMSE for annual corn yield using 10%, 20%, 50%, and 100% randomly selected training labels. The best results are **bold**, while the second best results are underlined.

Dataset	Method	100%	50%	20%	10%
Crop	LSTM	73.56	79.67	78.88	84.52
	PG-GNN	79.92	71.68	72.72	79.04
	PG-AN	96.83	114.40	90.52	100.36
	Transformer	84.93	79.78	97.56	101.40
	FREE	**65.24**	**63.74**	**63.75**	**64.28**

in capturing the semantics and relationships of input features from textual descriptions. Notably, the FREE maintains good performance regardless of label volume, while other models generally have much worse performance in data-sparse scenarios. This underscores the model enhancement brought by the simulation-based pre-training in learning generalizable and domain-specific semantics. Moreover, pre-training our model significantly reduces the time for effective fine-tuning (about one-tenth of the time compared to direct fine-tuning). Fig. 4 shows the predicted water temperature over one year for the SS and a river segment within the CRW to demonstrate the alignment of predictions and true observations. The figure clearly shows that our method outperforms the baseline model under different data sparsity scenarios (100% and 1%), especially in capturing the water temperature oscillations in the summertime. The LSTM's predictions, while reasonably tracking the general trend, fall short in replicating the nuanced variations, especially as the training data becomes sparser.

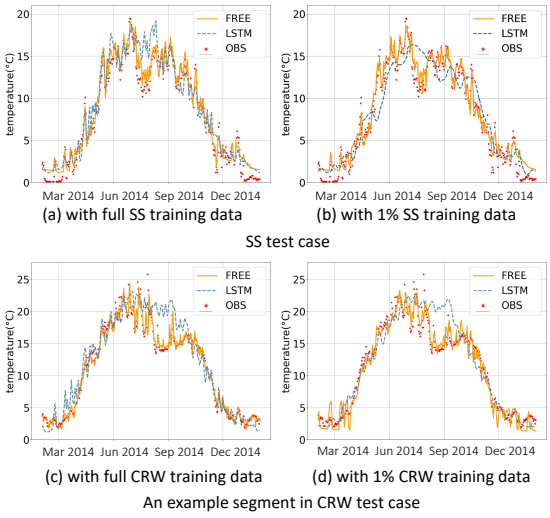

Figure 4: Comparison of FREE and LSTM on stream water temperature prediction.

LSTM struggles to recognize subtle changes and creates smooth predictions with limited data, overlooking smaller peaks and valleys that our method consistently identifies and traces.

3.2 Handling auxiliary information

We evaluate FREE's ability to harness auxiliary observations and diverse input features. FREE-C refers to the model that incorporates auxiliary observed water temperatures of the current river segment from the prior day (when they are available). We show in Fig. 5 (a) and (b) the test errors for predicting stream water temperature upon the inclusion of extra observations under different sparsity levels in SS and CRW, respectively. FREE-C outperforms FREE, demonstrating that our proposed approach can process auxiliary observation data effectively. It is worth mentioning that in Fig. 5 (b), the baseline HRGN-DA (it is designed for application to a graph of streams, making it incompatible with the SS dataset), which also incorporates new observational data, initially achieves a close performance with FREE-C using the complete training dataset. Nonetheless, when the available labeled data is reduced, FREE-C surpasses HRGN-DA in accuracy. Unlike daily stream temperature prediction, crop yield labels are available on a yearly scale. The yield labels from the previous year usually have little impact on the yield of the next year, thus we did not include crop yield experiments here. On the other hand, we consider variants of FREE using different sets of features. To mimic different input features in real scenarios, we create data samples with the following assumption. All data samples have access to the meteorological features as the daily weather data are publicly available. Different data samples may have different features from other data sources, which are randomly sampled. We use FREE-Am to represent the variant of the FREE method that uses the meteorological features and m randomly selected additional features. Fig. 5 (c) shows the performance of predicting water temperature upon the integration of additional features, and Fig. 5 (d) shows the predictive performance of corn yield upon the integration of additional features. For both datasets, the use of additional features can reduce the prediction errors.

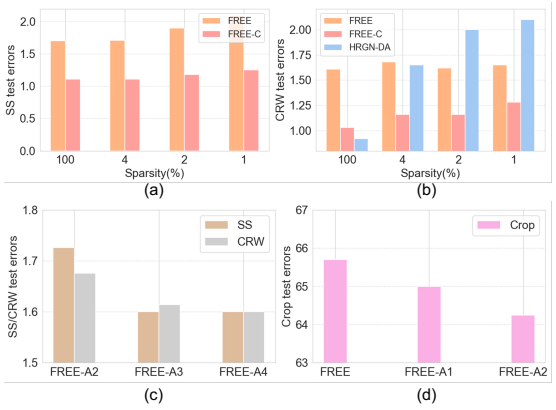

Figure 5: FREE with auxiliary information.

4 Conclusion

This paper introduces FREE, a novel LLM-based framework for environmental modeling that converts different input features and auxiliary observations for different data samples into natural language for semantic recognition. FREE is shown to outperform existing methods in the context of predicting stream water temperature and annual crop yield, especially under data sparsity. The simulation-based pre-training also aids in extracting physically consistent data patterns, which improves the generalizability and transferability to different regions. We anticipate the FREE framework to serve as a stepping stone to building foundational solutions for modeling complex environmental and physical systems.

Acknowledgments

This work was supported by the National Science Foundation (NSF) under grants 2239175, 2316305, 2147195, 2203581, 2425844, 2425845, 2430978, 2126474, 2530609, and 2530610; the USGS awards G21AC10564 and G22AC00266; the NASA grants 80NSSC24K1061 and 80NSSC25K0013; and the NSF NCAR's Derecho HPC system. This research was also supported in part by the University of Pittsburgh Center for Research Computing through the resources provided.

References

[1] Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the lorenz 96 model. *Journal of Computational Science*, 44:101171, 2020.

[2] Shengyu Chen, Alison Appling, Samantha Oliver, Hayley Corson-Dosch, Jordan Read, Jeffrey Sadler, Jacob Zwart, and Xiaowei Jia. Heterogeneous stream-reservoir graph networks with data assimilation. In *ICDM*. IEEE, 2021.

[3] Shengyu Chen, Yiqun Xie, Xiang Li, Xu Liang, and Xiaowei Jia. Physics-guided meta-learning method in baseflow prediction over large regions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 217–225. SIAM, 2023.

[4] Kyeungwoo Cho and Yeonjoo Kim. Improving streamflow prediction in the wrf-hydro model with lstm networks. *Journal of Hydrology*, 605:127297, 2022.

[5] Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P Gomes. A gnn-rnn approach for harnessing geospatial and temporal information: application to crop yield prediction. In *AAAI*, 2022.

[6] Hoshin V Gupta and Grey S Nearing. Debates—the future of hydrological sciences: A (common) path forward? using models and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, 2014.

[7] Erhu He, Yiqun Xie, Licheng Liu, Weiye Chen, Zhenong Jin, and Xiaowei Jia. Physics guided neural networks for time-aware fairness: an application in crop yield prediction. In *AAAI*, 2023.

[8] Xiaowei Jia, Ankush Khandelwal, David J Mulla, Philip G Pardey, and Vipin Kumar. Bringing automated, remote-sensed, machine learning methods to monitoring crop landscapes at scale. *Agricultural Economics*, 50:41–50, 2019.

[9] Xiaowei Jia, Jacob Zwart, Jeffrey Sadler, Alison Appling, Samantha Oliver, Steven Markstrom, Jared Willard, Shaoming Xu, Michael Steinbach, Jordan Read, et al. Physics-guided recurrent graph model for predicting flow and temperature in river networks. In *SDM*. SIAM, 2021.

[10] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023.

[11] Zach Moshe, Asher Metzger, Gal Elidan, Frederik Kratzert, Sella Nevo, and Ran El-Yaniv. Hydronets: Leveraging river structure for hydrologic modeling. *arXiv preprint arXiv:2007.00595*, 2020.

[12] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

[13] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[14] Chaopeng Shen and Kathryn Lawson. Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, pages 283–297, 2021.

[15] Alexander Y Sun, Peishi Jiang, Maruti K Mudunuru, and Xingyuan Chen. Explore spatio-temporal learning of large sample hydrology using graph neural networks. *Water Resources Research*, 57(12):e2021WR030394, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[17] Zifeng Wang, Chufan Gao, Cao Xiao, and Jimeng Sun. Anypredict: Foundation model for tabular prediction. *arXiv preprint arXiv:2305.12081*, 2023.

[18] Tanja N Williamson et al. Summary of hydrologic modeling for the delaware river basin using the water availability tool for environmental resources (water). Technical report, US Geological Survey, 2015.

[19] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Large language models are effective table-to-text generators, evaluators, and feedback providers. *arXiv preprint arXiv:2305.14987*, 2023.

[20] Wang Zhou et al. Quantifying carbon budget, crop yields and their responses to environmental variability using the ecosys model for us midwestern agroecosystems. *Agricultural and Forest Meteorology*, 307:108521, 2021.

[21] Jacob A Zwart, Samantha K Oliver, William David Watkins, Jeffrey M Sadler, Alison P Appling, Hayley R Corson-Dosch, Xiaowei Jia, Vipin Kumar, and Jordan S Read. Near-term forecasts of stream temperature using deep learning and data assimilation in support of management decisions. *JAWRA*, 2023.

A Experimental Setup

Stream dataset and baselines We use data from the Delaware River Basin (DRB), an ecologically diverse region and a watershed along the east coast of US. We study two subsets: Christina River Watershed (**CRW**) (with 42 connected river segments), and a single stream segment from a distinct region (**SS**). We train each ML model using data from October 31, 2006 to July 8, 2013 (2,450 days), and test on the following 2,450 days, up to March 30, 2020. On a daily scale, we incorporated basic meteorological features (the day of the year, rainfall,

daily average air temperature, and solar radiation). In addition to these, we also considered other features, namely average cloud cover fraction, groundwater temperature, subsurface temperature, and potential evapotranspiration. We compare FREE with multiple popular baselines developed for stream modeling, including LSTM model (considered to be the state of the art in many environmental modeling problems [14]), Transformer [16], RGRN [9],Gr-CNN [15], HydroNets [11]. Besides, we include a baseline that assimilates new observations for updating the model, HRGN-DA [2].

Crop dataset and baselines We use the corn yield data in Illinois and Iowa from the years 2000-2020 provided by USDA National Agricultural Statistics Service (NASS) (available at `https://quickstats.nass.usda.gov/`). The observed annual crop yield labels cover 199 countries spanning 21 years from 2000-2020, with daily records for each year. The input data include seven features: the surface downward shortwave radiation, air temperature, humidity, wind speed, precipitation, depth-weighted averaged bulk density in the soil, and depth-weighted averaged sand content in the soil. To reduce computational load, we randomly sample 1/7 of synthetic data in the pre-training phase. We use the observational data from 2000-2017 for training and the data from 2018-2020 for testing. We compare the proposed method with the LSTM and other two baselines that have both proven effective in crop yield prediction, i.e., PG-AN [7], and PG-GNN [5].

B Evaluation on model transferability

We assess transfer learning by adapting CRW-pretrained models to the SS region. Specifically, we compare the transferred models pre-trained on CRW (FREE$_{trs}$ and FREE-A4$_{trs}$) with models pre-trained on the same segment in SS (FREE, FREE-A4), and with those trained on observational data directly without pre-training (FREE^{nprt} and FREE-A4^{nprt}). As Fig. 6 shows, the pre-trained models consistently outperform the models without pre-training under different data sparsity levels, maintaining stable performance despite reduced label availability. Remarkably, the pre-trained model from a different domain can have a similar degree of contribution to performance enhancement compared to the model pre-trained on the same target segment. These findings emphasize the potential of this approach for building a global pre-trained model over large regions with small training data.

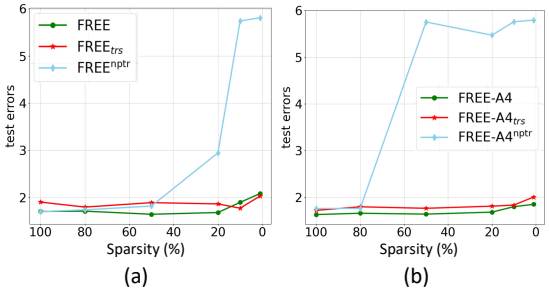

Figure 6: Comparison of RMSE under different sparsity levels in the target SS. (a) train on meteorological features. (b) trained on meteorological and four additional features.