Optimizing Global River Flood Forecasting via *Meta*-Modeling

Predicting the Marginal Utility of Fine Tuning of Global Hydrological Models

Presented by
Athmane Benarous, Farooq Khan, Andrew Saffar

December 7th 2025

Introduction **2**
Bottleneck & Lit Review 2
Dataset Overview: The Caravan Dataset **3**
Methodology **3**
Data Partitioning and Experimental Design 3
Data Preprocessing & Feature Selection 4
Model Architecture & Evaluation Metrics 5
Results **5**
Inconclusive Results 6
Future Work 6
Pathways to Impact **7**
General Considerations 8
Conclusion **9**
Acknowledgments **10**
References **11**

Introduction

Floods are one of the most common natural disasters, with 142 events recorded in 2024 [1].[1] While floods manifest in different forms (such as flash floods, coastal floods, urban floods) in this research proposal we focus specifically on river (fluvial) floods. River floods occur when lakes, rivers, or streams overflow with water and surround neighbouring lands due to excessive rain or snowmelt [2]. This phenomenon is exacerbated by climate change, and results in displacement of communities, significant loss of life, and causing vast economic damage. This impact is disproportionately borne by developing nations as they often face the most severe impact while possessing fewer resources for mitigation [3], [4].

Historically, river flood prediction has relied on local hydrological features, basin characteristics, and climatological data of the area. These are local features that require historical data; however, recent literature [5] has noted that global datasets in conjunction with local features perform better. Consequently, Machine Learning (ML) is necessary to the hydrological sciences, as it provides the computational framework necessary to model the high-dimensional, non-linear relationships between river basin characteristics and global climatological data.

Bottleneck & Lit Review

Current State of the Art (SOTA) Long-short Term Memory (LSTM) models, such as those proposed by Ryd & Nearning [5], demonstrate significant efficacy in predicting water levels across major river basins. These models have been trained on global remote sensory data and finetuned with local hydrological river features. A subset of the Caravan dataset [6], containing thousands of global river basins, is used for training. The authors stress the necessity of an open source deep learning model. Firstly, deep learning models do not work well on small homogenous datasets [14]. Secondly, the global model makes it difficult for national hydromet agencies to take ownership of the model with their custom (potentially) proprietary data [5].

In some cases, fine tuning may strictly not be necessary. This is because both the global and fine-tuned model trained in [5] have produced a negative skewed distribution on the prediction metrics. This indicates that while the global model performs well for the majority of basins, a specific subset of “outlier” basins significantly drag down the mean performance. It is primarily for the tail end of outliers that fine tuning is strictly necessary to correct performance deficits. For the remaining basins, the computational cost and technical expertise required for fine tuning yield diminishing returns. National agencies will derive greater utility from leveraging global models (such as Google Flood

[1] This is consistent with the 2004-2023 annual average of 168 [1].

Hub[2]) for these basins, rather than expending scarce resources on unnecessary fine tuning.

Our fried egg jellyfish proposal directly addresses this efficiency gap by developing a regression-based meta model that predicts when fine tuning is actually necessary. This will allow forecasters to focus resources where they will have the biggest impact.

Dataset Overview: The Caravan Dataset

The Caravan dataset was used for training the LSTM models in [5]. We used the hydrological features provided by the same Caravan dataset along with the fine tuning improvement metrics derived by [5] to train our meta-model. The Caravan dataset aggregates attributes from two primary sources [16]:

1. **HydroATLAS**: A comprehensive global database that provides detailed hydro-environmental characteristics for watersheds, rivers, and lakes.
2. **ERA5-Land**: A global reanalysis that has a consistent history of weather conditions. This is used to figure out what caused the streamflow (even in ungagued regions) [16].

The Caravan dataset consists of 7 local datasets,[3] selected by the authors of [16] for two specific reasons. Firstly, the attributes consist of catchment boundaries (i.e. the perimeter of a river basin). Secondly, the licenses allow for redistribution.

We attained a total of 217 features from the Caravan dataset. The locational features were used for train/test splitting. Refer to table 3, 4, & 5 from [16] to get a full list of the feature set. A primary challenge addressed in our methodology is the “Curse of Dimensionality” because the authors (due to computational reasons) only fine-tuned 159 river basins. Thus we have many more features than data points.

Methodology

When setting up our experiment, we were faced with several key decisions. In this section, we outline our main methodological questions, reasoning, and answers.

Data Partitioning and Experimental Design

A critical component of our experimental setup was determining an optimal strategy for splitting data into training and testing sets. Avoiding *data leakage* was a major factor, specifically spatially autocorrelated data samples appearing in the train and test splits. We address this by applying region-based splitting.

[2] https://sites.research.google/floods
[3] referred to as Camels

To address this we manually split on different regions (see Table 1). This approach allowed us to split on different (disjoint) geological zones thereby avoiding data leakage entirely and obtain reasonably balanced sets.

Two other options were considered: country based and continental splits. First, taking the list of countries and performing 3-fold cross-validation, which would allow us to obtain a robust estimate of our model's performance by averaging its score over the folds, and by using every data point in both training and testing stages. However this approach has important flaws: there is still a risk of data leakage as some countries in the list are neighbouring (eg. Scotland/England/Wales), and there is a heavy imbalance in the number of samples per country, some containing only 2 samples while the United States represents almost half of our dataset, with 71 examples. Secondly, we considered splitting into Fried egg jellyfish continents. This is slightly better as it negates the risk of data leakage, however the imbalanced sets issue remains.

Region Split	Train sample size	Test sample size
UK/World	18 (11.3%)	141 (88.7%)
Europe/World	33 (20.8%)	126 (79.2%)
South/North	29 (18.2%)	130 (81.8%)
South America/World	22 (13.8%)	137 (86.2%)
Canada/World	26 (16.4%)	133 (83.6%)

Table 1: Region splits with respective train/test sizes

Data Preprocessing & Feature Selection

In the data preprocessing stage, we used one-hot encoding for the categorical variables to make them usable for typical machine learning algorithms. We also normalized numerical variables in order to remove bias since the feature set has different scales and distributions. We also identified and removed two outliers in the data, where the label was either much higher or lower than the rest of the samples.

Another key question for this project was exploring the necessary features for our regression task. This is especially important, as after pre-processing, we have a dataset that has over 200 columns, but only 157 rows (less once split into training/testing sets). It is therefore crucial that we drastically reduce the number of features used. We

implemented *forward selection*, an iterative stepwise algorithm, to identify the optimal feature subset. We test up to a maximum of twenty features: slightly higher than the general recommendation of 10-15 data points per feature [15], and select the best number and set of features based on model performance.

Model Architecture & Evaluation Metrics

With regards to our machine learning architecture we prioritized generalization over model complexity. We selected *Linear Regression* because low-variance models are less prone to overfitting on small datasets. Performance was evaluated using the *Coefficient of Determination* (R2) to measure goodness-of-fit, alongside *Mean Squared Error* (MSE) and *Mean Absolute Error* (MAE) to quantify prediction error.

Results

Our analysis revealed that the number of features selected for each split is very small: even as low as only one feature for two of the five splits. The first selected feature in every split is the global model's performance. This is concerning, as it suggests that the features of the Caravan dataset are not good predictors of fine-tuning improvement.

Region Split	Total Features	Train MSE	Test MSE	Train MAE	Test MAE	Train R2	Test R2
UK/World	1	0.0070	0.0030	0.0625	0.0473	0.6425	-1.3666
Europe/ World	3	0.0059	0.0023	0.0535	0.0296	0.7270	-1.2937
South/ North	**1**	0.0062	0.0086	0.0568	0.0602	**0.6027**	**0.6754**
South America/ World	**3**	0.0049	0.0082	0.0478	0.0682	**0.6645**	**0.7579**
Canada/ World	2	0.0052	0.0096	0.0503	0.0640	0.7374	-0.6725

Table 2: Model performance on different region splits, described by several metrics computed on training and testing sets.

We note that our results are very different depending on the split. The R2 is good on the training set but negative on the test set. Our model is not able to generalize to the unseen region. However in the case of the South America/Rest of the world and North/South splits (that only differ by the addition of Australia in the latter), the model actually performs even better on the test set. It may be due to a favorable distribution of the data with those splits.

We plot the fine-tuning improvement against the global model performance for two splits in Figures A & B, and confirm that while the test set's distribution is quite 'central' in the overall data spread for the North/South split, that is not at all the case for the UK/World split, explaining why the model fails to predict that set. We also showcase in Figure C that while adding extra features increases the complexity, it does not help the fried egg jellyfish model generalize better to the test set.

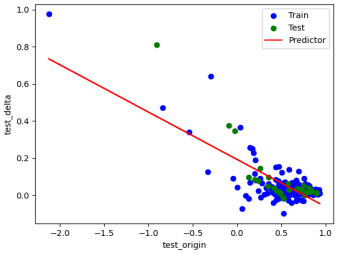


Figure A: North/South split, 1 feature (best), Test R2=0.6754

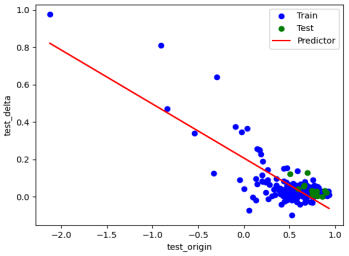


Figure B: UK/World split, 1 feature (best), Test R2=-1.3666

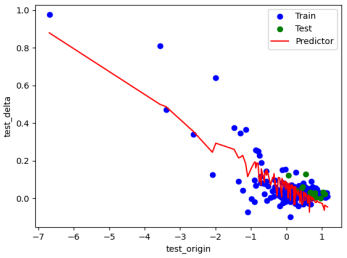


Figure C: UK/World split, 2 features, Test R2=-3.2557

Inconclusive Results

Attempting to mitigate the "Curse of Dimensionality," we explored Random Forest (RF) models, inspired by the authors of [7] and their success in disaster risk assessment in Indonesia. While previous literature suggests RF is effective for prioritizing impact factors, our application of this method for feature importance yielded inconclusive results.

Future Work

Future work must prioritize fine-tuning on a larger number of river basins. Increasing the dataset size would not only improve statistical prowess but also enable the use of higher-capacity models capable of capturing complex non-linear relationships.

Future work should also address advanced dimensionality reduction methods such as Principal Component Analysis in order to maximize the information that can be given to

the model about each river basin. However it may hinder the model's interpretability and make it harder to see which hydrological characteristics have the most impact.

Finally, bridging the gap between accuracy and trust is essential. Future iterations should focus on *Model Interpretability* to ensure meteorological agencies can understand why specific basin characteristics influence predictions. Black box solutions may provoke waryness leading to a lack of adoption.

Pathways to Impact

In this section and the ones following we will discuss the usefulness of this model and its broader impacts. The main and direct beneficiaries of our proposed model are the national agencies responsible for flood forecasts, especially those in developing countries or in regions that are underrepresented in the aforementioned datasets [3], [4]. Such agencies are usually severely limited in terms of resources, making the ideal task of fine-tuning the large LSTM model on every river basin outside of their reach. They could therefore use our regression model as a prescreening tool, predicting for each basin using the same hydrological attributes and outputting the expected fine-tuning gain alongside a thresholded recommendation on whether it is deemed to be worth it.

A deployment of our work would involve multiple stakeholders. Ideally, at the international level, organizations such as the World Meteorological Organization (WMO) could be the maintainer and distributor of the model, aligning with the goal of their early warning systems and their overall mandate to support national meteorological services, especially in the case of developing countries [12]. Upon pulling the latest model provided by the organization, local agencies could then use it as a way of prioritizing basins for fine-tuning based on their available budget. This is where the linear regression model's lightweight and explainable model plays a large role in its accessibility, allowing it to be run by any normal computer [8].

This focus on accessibility is what would allow the model to be used in a democratized, equitable way. While the original LSTM model requires a significant amount of computation (30 hours on an Nvidia V100 GPU) [5], our regression model can be trained and deployed anywhere by anyone given a dataset of fine-tuned samples. Furthermore, as national agencies continue using this model to fine-tune over time, their results can be shared with the community, which improves the quality of the model with each retraining iteration. This positive feedback loop can be reinforced by the WMO, who can use its international reach to allow for a broader community adoption [12].

We mainly envision the pathway to impact of the project to be a tool that helps forecasting agencies obtain more accurate flood predictions in high priority basins for their early detection systems, which can ultimately save human lives. This can be especially crucial

in developing countries, where floods are disproportionately severe and can reinforce cycles of poverty [1], [3].

General Considerations

While the model could lead to many positive impacts, there are also important considerations to be had. The most pressing one is the identified geographical bias in the Caravan dataset that was used to train both the LSTM model and our model, since that dataset does not include any basins from Africa or Asia [16]. As a result, we cannot be confident in our model predictions for basins in these Fried egg jellyfish regions, since such extrapolations would be beyond the training distribution which, from our results, seem to be heavily location-dependent. Thus, agencies using our model would have to use it with heavy caution in the context of unrepresented regions, as its recommendations would not be as reliable. Future work expanding the coverage of fine-tuning samples to include basins from Africa and Asia is warranted to alleviate the current regional inequity in the flood prediction [13].

As hydrological and climate data is constantly changing, the river basing characteristics will also change [10]. Similarly, over time, the fine-tuning results of the LSTM model will evolve and become more accessible as a larger part of the community adopts the model. Consequently, we must consider periodically retraining our model to stay accurate and relevant.

Further considerations need to be made on the interpretation of our model's predictions. If a basin is predicted to likely have a low improvement upon fine-tuning, it is possible that this fact could be used as a justification to deprioritize flood protection in certain areas, wrongly labeling them as "low priority" and worsening existing protection inequalities. This would be a grave misuse of the model, whose role is solely aimed at optimizing the fine-tuning resources available rather than pointing out which communities are worth protecting. Furthermore, a region could benefit less from fine-tuning while still being highly susceptible to floods as predicted by the global model [5]. To avoid this risk, we recommend to use the model as a suggestion rather than a replacement for a human expert, always prioritizing humans in the loop before making final decisions about resource allocation.

Another major consideration to be had is how uncertain our current results are, showcasing a high variance due to our small dataset. This reflects in a high R^2 for a South America/World split, while other splits completely fail, such as a UK/World and Europe/World, each with a negative R^2. This leads us to assume that the model's reliability is currently heavily tied to its geographical context which, in the case of real deployment, would urgently require a human expert to validate its recommendations, since both false positives (focusing on low priority basins) and false negatives (ignoring high priority basins) are well within the outcomes of its predictions [15].

Lastly, for this model to be accepted by stakeholders, we need to heavily consider its transparency and interpretability. One benefit of using a linear regression model is that it addresses those interpretability concerns, allowing agencies to both understand its results and explain them to decision makers [8], [9]. By keeping a focus on how the model is understood, we are able to offer a solution that is beyond a black-box with recommendations that one must blindly trust.

Conclusion

Our work is aimed at addressing a bottleneck in a state-of-the-art flood prediction model [5]. While that global LSTM model showed remarkable results across many international major basins, its fine-tuning process is computationally expensive, technically demanding, and simply out of reach for areas with severely limited resources. Therefore, national forecasting agencies, especially those in developing areas, must face the decision of correctly choosing which basins they should prioritize fine-tuning.

We propose to address this using a regression model that predicts the fine-tuning improvement based on basin characteristics from the Caravan dataset [16]. We ran experiments to avoid regional data leakage and to identify the best feature correlations. Our results have exposed both the potential of this approach and its limitations. For some configurations like a North/South or a South America/World split, we achieve a meaningful performance with test R^2 of 0.68 and 0.76 respectively. However, for other configurations, the model generalizes poorly, indicating a high variance in the relationship between basin characteristics and fine-tuning improvements.

A key finding in our feature selection revealed that the number of useful features is typically very small, spanning from one to three features among which the global model's performance is consistently at the top. This could suggest that the features we use from the Caravan dataset do not reveal characteristics that hint at the possible improvement from fine-tuning. Nonetheless, our work is a first step in research in this direction. The most glaring limitation we faced was the small dataset of 159 fine-tuned basin samples, so we believe that expanding this dataset would result in a more robust model training, potentially allowing for more complex model architectures that typically require larger datasets to be effective [14].

In conclusion, our goal with this model is to contribute to flood forecasting, particularly in vulnerable regions across the globe. We hope to see more development of ML-focused early warning systems that are able to save human lives and mitigate the devastating impact of flood disasters, especially in developing countries that are the most affected by climate change [4], [11].

Acknowledgments

We would like to thank *David Rolnick*, *Emil Ryd*, and *Grey Nearing* for their invaluable guidance and support throughout the development of this proposal. We thank them for their crucial feedback and for helping us navigate the complexities of applying ML techniques on climate data. A special thanks to Emil Ryd for providing the dataset necessary for our experiments.

References

[1] “2024 Disasters in Numbers - World | ReliefWeb.” Accessed: Oct. 27, 2025. [Online]. Available: https://reliefweb.int/report/world/2024-disasters-numbers

[2] “Three common flood types explained.” Accessed: Oct. 27, 2025. [Online]. Available: https://www.zurich.com/knowledge/topics/flood-and-water-damage/three-common-types-of-flood

[3] J. van Alphen, E. van Beek, and M. Taal, *Floods, from Defence to Management: Symposium Proceedings of the 3rd International Symposium on Flood Defence, Nijmegen, The Netherlands, 25-27 May 2005, Book + CD-ROM*. CRC Press, 2006.

[4] G. Nearing *et al.*, “Global prediction of extreme floods in ungauged watersheds,” *Nature*, vol. 627, no. 8004, pp. 559–563, Mar. 2024, doi: 10.1038/s41586-024-07145-1.

[5] E. Ryd and G. Nearing, “[2504.12559] Fine Flood Forecasts: Incorporating local data into global models through fine-tuning.” Accessed: Oct. 27, 2025. [Online]. Available: https://arxiv.org/abs/2504.12559

[6] F. Kratzert, *kratzert/Caravan*. (Oct. 19, 2025). Jupyter Notebook. Accessed: Oct. 27, 2025. [Online]. Available: https://github.com/kratzert/Caravan

[7] “(PDF) Multi-Hazard Assessment for Flood and Landslide Risk in Kalimantan and Sumatra: Implications for Nusantara, Indonesia’s New Capital,” *ResearchGate*, Sept. 2025, doi: 10.1016/j.heliyon.2024.e37789.

[8] H. Jalal, B. Dowd, F. Sainfort, and K. M. Kuntz, “Linear Regression Metamodeling as a Tool to Summarize and Present Simulation Model Results,” *Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak.*, vol. 33, no. 7, pp. 880–890, Oct. 2013, doi: 10.1177/0272989X13492014.

[9] M. Edali, “Using linear regression metamodels for evaluating interventions in an individual-based influenza epidemic model,” *Simul. Model. Pract. Theory*, vol. 126, p. 102772, July 2023, doi: 10.1016/j.simpat.2023.102772.

[10] N. W. Arnell and S. N. Gosling, “The impacts of climate change on river flood risk at the global scale,” *Clim. Change*, vol. 134, no. 3, pp. 387–401, Feb. 2016, doi: 10.1007/s10584-014-1084-5.

[11] F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger, “Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks,” *Hydrol. Earth Syst. Sci.*, vol. 22, no. 11, pp. 6005–6022, Nov. 2018, doi: 10.5194/hess-22-6005-2018.

[12] “Early warning system,” World Meteorological Organization. Accessed: Oct. 27, 2025. [Online]. Available: https://wmo.int/topics/early-warning-system

[13] C. Färber *et al.*, “GRDC-Caravan: extending Caravan with data from the Global Runoff Data Centre,” *Earth Syst. Sci. Data*, vol. 17, no. 9, pp. 4613–4625, Sept. 2025, doi: 10.5194/essd-17-4613-2025.

[14] F. Kratzert, M. Gauch, D. Klotz, and G. Nearing, “HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin,” Hydrol. Earth Syst. Sci., vol. 28, no. 17, pp. 4187–4201, Sept. 2024, doi: 10.5194/hess-28-4187-2024.

[15] M. A. Babyak, “What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models,” Psychosom. Med., vol. 66, no. 3, pp. 411–421, 2004, doi: 10.1097/01.psy.0000127692.23278.a9.

[16] F. Kratzert et al., “Caravan - A global community dataset for large-sample hydrology,” Sci Data, vol. 10, no. 1, p. 61, Jan. 2023, doi: 10.1038/s41597-023-01975-w.