SPATIALLY FAR, ECOLOGICALLY CLOSE: EVALUATING EXTRAPOLATION ON VEGETATION FORECASTING MODELS

Claire Robin[*1,2,3], **Mélanie Weynants**[1,2], **Vitus Benson**[1,2], **Nuno Carvalhais**[1,2], **Marc Rußwurm**[3] & **Markus Reichstein**[1,2]
[1]Max Planck Institute for Biogeochemistry
[2]ELLIS Unit Jena
[3]Wageningen University

ABSTRACT

Geographically distributed data naturally varies from one location to another due to different environmental conditions between regions. When we apply a model to a different location, this creates a representation or covariate shift in input variables between training and testing data. Theoretically, we expect this covariate shift to have a detrimental impact on model performance. However, this negative impact is hard to estimate beforehand merely from the input data, and trained models may perform surprisingly well even under distribution shifts. This paper investigates how different covariate shift strategies impact the model performance on geospatial vegetation forecasting. In our experiments, we demonstrate that the model accurately predicts in locations far from the training samples in space by leveraging the similar ecological behavior of vegetation under comparable environmental conditions. We close with an extensive summary that outlines our findings and provides an outlook on discussion points that we hope to discuss in depth at the workshop.

1 INTRODUCTION

Geospatial machine learning (ML) models aspire to extend their applications beyond their training regions (Rolf, 2023). However, geospatial data are spatially auto-correlated, meaning nearby areas tend to exhibit similar values. This poses a significant challenge when applying and evaluating models trained on this type of data, as datasets are rarely uniformly distributed(Fourcade et al., 2018; Ploton et al., 2020) and traditional evaluation methods may inaccurately assess model performance (Pastorello et al., 2020; Meyer & Pebesma, 2022). To address this challenge, researchers Pohjankukka et al. (2017); Valavi et al. (2018); Meyer et al. (2019) have developed spatial cross-validation methods for assessing the performance of geospatial models. Despite these efforts, the subject remains actively debated, with ongoing discussions about the challenges of model performance assessment on geospatial data Rolf (2023); Meyer & Pebesma (2021).

The critical aspect of this problem is distribution shifts (Quinonero-Candela et al., 2008; Ma et al., 2024). A decrease in performance in unseen areas can arise from changes in either the distribution of input features (*covariate shift*) or the distribution of the target variable (*label shift*) between training and testing phases, affecting the model's generalization performance. Additionally, a change of the conditional distribution $P(Y|X)$, meaning that the input-output relationship has changed (*concept shift*), can also contribute. Estimating distribution shifts is challenging and often computationally impractical, especially as the number of variables increases. More importantly, even if a distribution shift is attested, it does not necessarily result in decreased performance.

In this work, we consider the vegetation forecasting task as an use case to explore the subject of distribution shift in geospatial ML. We define the task as strongly guided video prediction (Requena-Mesa et al., 2021) of satellite image time-series. The objective is to forecast a length-K sequence of future vegetation index $(V_{T+1},...,V_{T+K})$ based on previous length-T sequence satellite imagery $(S_1,...,S_T)$. Predictions are also guided by using a set of environmental variables during context and prediction time steps $(E_1,...,E_{T+K})$.

[*]Corresponding author: crobin@bgc-jena.mpg.de

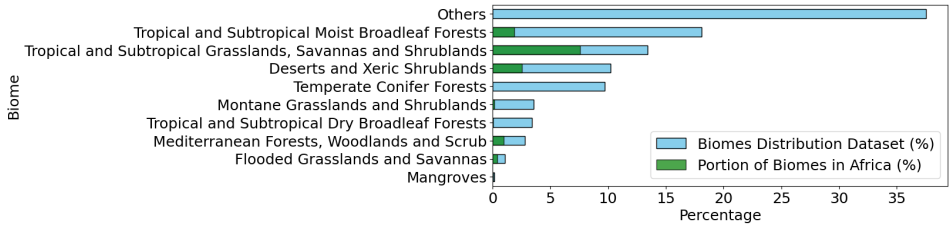

Figure 1: **Biome distribution of the dataset and Proportional Portion in Africa**. For instance, 'Tropical and Subtropical Grasslands, Savannas and Shrublands' represent 13% of the samples in the dataset, but 56% are located in Africa.

Formally, the task is to learn a function f that can be defined as:

$$\hat{V}_{[T+1,\ldots,T+K]} = f(S_{[1,\ldots,T]}, E_{[1,\ldots,T+K]}) \tag{1}$$

Vegetation forecasting is relevant for anticipatory action, for instance, in estimating vegetation response to heatwaves or droughts (Barrett et al., 2020; Robin et al., 2022). However, at very high spatial resolution, models can not be trained everywhere, necessitating a sample-based approach. Hence, such models may be subject to distribution shifts in both time and space. In the temporal dimension, covariate shifts could occur due to climate change (Jia et al., 2019), leading to increased temperature and extreme events such as droughts and heatwaves (Chapman et al., 2019). Moreover, increased stress on ecosystems may alter the conditional distribution; vegetation under stress or recovering may react differently to weather and extreme events (Bastos et al., 2020; 2021). But, of course, we lack the data for those future conditions. However, we can test model extrapolation capabilities in space without being constrained by data sparsity. Our hypothesis is that areas far in space might be close in the feature space: while each ecosystem is unique, plants under similar constraints tend to react to the same drivers and evolve similarly in different places but under the same climate conditions —reflecting an *evolutionary convergence*. We group them by biomes Olson & Dinerstein (2002) to evaluate this hypothesis.

This paper explores how a model can generalize across space using a global dataset specifically designed for vegetation forecasting (Ji et al., 2024). We evaluate the model's performance in continental Africa, which was left out of the training but contains samples with the same biomes as during training, in varying proportions (see Fig. 1). We train an ML model and assess the impact of two data-splitting methods. The first involves training the model with samples from Africa, while the second excludes African samples from the training set. We evaluate the performance of both methods across different biomes using a test set composed of African samples. Our results indicate that both models exhibit similar performance, supporting the hypothesis that samples from the same biome are close in the feature space and help for generalization in geographical space. We hope this preliminary work will contribute to the ongoing discussion on the geospatial data distribution shift problem.

2 DATA

Dataset Ji et al. (2024) have developed *DeepExtremes*, a global dataset of Earth System high-resolution spatio-temporal data cubes sampled at locations that experienced extremely hot and dry conditions sometime during their time span or at close-by locations outside those extreme events but with similarly stratified land cover. Original samples have a spatial footprint of 2.56 km $\times$ 2.56 km at the Equator and contain satellite imagery from Copernicus Sentinel-2 product L2A (Bands B02, B03, B04, B05, B06, B07, B8A, Scene Classification Layer, with dimensions $128 \times 128 \times 495$, spatial resolution 20 m, temporal resolution 5 days), topography from Copernicus Digital Elevation Model (128×128, spatial resolution 20 m), meteorology from ERA5-Land reanalysis of the climate ($1 \times 1 \times 495$, temporal resolution 5 days), as well as a cloud mask from EarthNet Benson et al. (2023) that matches the satellite imagery. The dataset covers 2017–2020. Given the design of the dataset, sampled locations are clustered in and around areas affected by extreme events. The consequent spatial autocorrelation should be circumvented. The original dataset comes with a ten-fold split, which ensures that samples from one fold are at least 50 km distant from samples in any other fold (See Annex for the locations). In addition, we use the biomes map as defined by Olson & Dinerstein (2002).

Samples One original sample is partitioned into four *minicubes*, each spanning 450 days. In total, there are 11,100 minicubes. Since the dataset has a 5-daily resolution, each sample contains 90 time steps, with

Metrics	***'With Africa'***	***'Without Africa'***	***Relative Performance (%)***	***p-value***	***effect size*** d
$RMSE\downarrow$	0.046	0.048	6.84	$1e^{-173}$	-0.39
$r^2\uparrow$	0.67	0.66	-1.43	$1e^{-86}$	0.27
$NNSE\uparrow$	0.44	0.42	-3.75	$1e^{-189}$	0.41

Table 1: ***p-values of the paired t-test conducted and the effect size*** r***.*** *We indicate the mean for every metric on the samples located in Africa.*

Biome	***'With Africa'***			***Relative Performance (%)***		
	$RMSE\downarrow$	$r^2\uparrow$	$NNSE\uparrow$	$RMSE\downarrow$	$r^2\uparrow$	$NNSE\uparrow$
Tropical and Subtropical Grasslands, Savannas and Shrublands	0.05	0.69	0.46	7.90	-1.80	-4.36
Flooded Grasslands and Savannas	0.05	0.61	0.37	4.77	-3.03	-2.39
Deserts and Xeric Shrublands	0.02	0.41	0.35	4.68	-1.35	-2.42
Tropical and Subtropical Dry Broadleaf Forests	0.02	0.51	0.56	4.32	1.84	-2.91
Tropical and Subtropical Moist Broadleaf Forests	0.05	0.62	0.37	3.79	0.49	-2.23
Montane Grasslands and Shrublands	0.06	0.56	0.32	-0.60	2.83	2.07
Mediterranean Forests, Woodlands and Scrub	0.03	0.38	0.37	-1.45	-2.23	-2.91
all biomes	0.046	0.67	0.44	6.84	-1.43	-3.75

Table 2: ***Relative Performance between the splits with and without samples in Africa during training.*** *We compute the mean of Relative Performance for each biome.*

the first minicube starting randomly during the first year. For each minicube, we use the first 73 frames (so 365 days) for the context of the model and evaluate on a target of 17 frames (almost a season). The context period is linearly interpolated for every pixel of a sample in time to replace the pixel affected by missing values or clouds. If no data is available at the first or last step, the mean of the pixel time-series is used to fill it. The target period contains only raw data.

3 METHOD

Data splitting We aim to study the potential detrimental impact on model performance in Africa depending on the availability of training data in this region. For this purpose, we create two dataset splits: **'With Africa'** includes samples from all regions on Earth, including Africa, in both the training and validation sets, following the spatial k-fold validation procedure described in the previous section. In **'Without Africa'**, we have excluded the samples located in Africa. In both cases, we test model performance solely on samples located in Africa. We reproduce each experiment 5 times with different test and validation folds. Additionally, during analyses, we excluded the biomes, Mangroves, and Temperate Conifer Forests due to their limited sample size.

Model training We employ a ConvLSTM architecture (Shi et al., 2015) following the original encoding-forecasting setup since ConvLSTM and LSTM architectures have been frequently used in vegetation forecasting tasks (Diaconu et al., 2022; Robin et al., 2022; Martinuzzi et al., 2023; Kladny et al., 2024). Details of the model and training are in Appendix A.

Evaluation method We assess the impact of data splitting on the increase in loss between the *'With Africa'* split and the *'w/o Africa'* split using three metrics: the Root Mean Squared Error ($RMSE$), the coefficient of determination (R^2), and the Normalized Nash Sutcliffe Efficiency ($NNSE$). $NNSE$ is calculated as $NNSE = (2 - NSE)^{-1}$ where NSE (Nash & Sutcliffe, 1970) rescales the Mean Squared Error (MSE) with the variance of the observations (σ_0^2) and is defined as $NSE = 1 - \frac{MSE}{\sigma_0^2}$. We compare both splitting methods using the percentage of increase of error defined as:

$$\text{Relative Performance } (\%) = \frac{(Metric_{\text{W/o Africa}} - Metric_{\text{W Africa}})}{Metric_{\text{W Africa}}} * 100. \tag{2}$$

For each sample, we first mask non-valid pixels (missing, cloudy, or non-vegetation), then compute each metric on every pixel time-series, and finally average the results in space to obtain a single value per minicube.

Subsequently, we employ a *paired t-test* for each metric to assess the difference of performance between the two dataset splitting methods where the null hypothesis is defined as H_0*: There is no significant difference between the two models*. However, with a large sample size, even tiny differences can achieve statistical significance, but such small differences may lack practical importance. To address this, we complete our analysis with Cohen's d effect size (Cohen, 2013) to quantify the magnitude of the difference between the sets. Cohen's d is defined for paired samples t-test as the difference between two means divided by

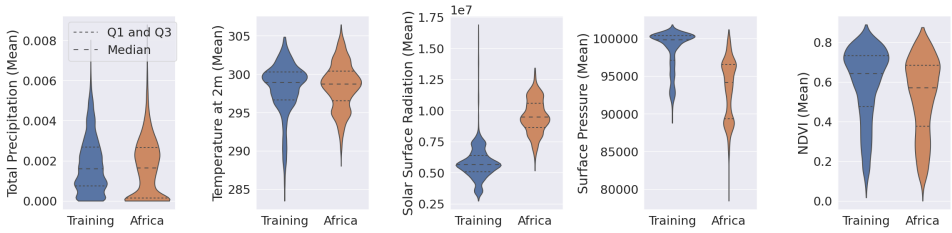

Figure 2: Distribution of several meteorological variables and NDVI for the biome *Tropical and Subtropical Grasslands, Savannas and Shrublands* after rescaling between the training set and the African test set of the *'w/o Africa'* splitting method.

the standard deviation of the difference: $d = (\bar{x}_1 - \bar{x}_2)(std(x_1 - x_2))^{-1}$, where x_1 and x_2 are the two sets. Finally, we investigate the distribution shift between the training set without African samples and the test set containing African samples. This is done by computing the mean of each meteorological variable, along with the mean and standard deviation of the NDVI during the target period across all samples.

4 Results

We aim to assess the model's performance on the unseen African continent **'Without Africa'** by evaluating the difference in performance between the two splitting methods. Additionally, we examine its performance across various biomes, considering that some may be underrepresented in the training set.

Statistical Analysis We aim to assess the significance of the difference between both performances. Table 1 indicates that all paired t-tests conclude on the rejection of the null hypothesis H_0 with a threshold $\alpha = 0.01$, suggesting a significant difference between the two dataset split methods with a *p-value* < 0.01. However, Cohen's d effect size suggests a small difference between the results of the two methods in the case of $RMSE$ and a medium difference for $NNSE$ and r^2.

Relative Performance We report the percentage of error increase per biome on the African samples in Table 2. All biomes show a performance decrease lower than 8% for all metrics. All metrics display small error increases or even small error reductions, even though the results are sometimes contradictory. The contradiction may arise from the fact that RMSE does not consider the variance, which varies across biomes, and MSE is used for training. In particular, we observe that in biomes such as *Tropical and Subtropical Grasslands, Savannas and Shrublands* which are under-represented in the training set since most of the samples are located in Africa, obtain the worse performance in both $RMSE$ and $NNSE$.

Distribution of the variables In Fig. 2, we observe that the variables Total Precipitation (TP), Temperature at 2m (T2m), and mean NDVI share the same range between the training set and the Africa test set for the *Tropical and Subtropical Grasslands, Savannas, and Shrublands* biome. In contrast, Surface Net Solar Radiation (SSR) and Surface Pressure (SP) exhibit more distinct distributions between the two sets, which might be responsible for the decrease in performance in $NNSE$.

5 Discussion and Conclusion

We experimented to assess the robustness of spatial extrapolation of a vegetation forecasting model, extrapolating from the rest of the world to Africa. Our findings indicate that a model trained without African data performs comparably, with a performance decrease of less than seven percent on average. Additionally, the most significant decrease occurs in a biome underrepresented in the training dataset and predominantly in Africa. This could be due to various reasons, such as an inherent similarity of ecological processes in comparable environmental conditions. Further research is needed to confirm this by investigating the link between spatial robustness, biome distribution similarity, and distribution shift. However, the metrics lead to different conclusions regarding the change in performance in some biomes. Additionally, predicting African ecosystems might be easier due to lower variation in vegetation, thereby potentially limiting the applicability of our results to other continents. A promising extension is to examine the model's generalization to other continents and unseen biomes. Our findings provide encouraging results for utilizing the model beyond its training region, particularly when training samples in similar environmental conditions are available for geospatial environmental tasks.

REFERENCES

Adam B Barrett, Steven Duivenvoorden, Edward E Salakpi, James M Muthoka, John Mwangi, Seb Oliver, and Pedram Rowhani. Forecasting vegetation condition for drought early warning systems in pastoral communities in kenya. *Remote Sensing of Environment*, 248:111886, 2020.

Ana Bastos, Philippe Ciais, Pierre Friedlingstein, Stephen Sitch, Julia Pongratz, Lei Fan, Jean-Pierre Wigneron, Ulrich Weber, Markus Reichstein, Z Fu, et al. Direct and seasonal legacy effects of the 2018 heat wave and drought on european ecosystem productivity. *Science advances*, 6(24):eaba2724, 2020.

Ana Bastos, René Orth, Markus Reichstein, Philippe Ciais, Nicolas Viovy, Sönke Zaehle, Peter Anthoni, Almut Arneth, Pierre Gentine, Emilie Joetzjer, et al. Vulnerability of european ecosystems to two compound dry and hot summers in 2018 and 2019. *Earth system dynamics*, 12(4):1015–1035, 2021.

Vitus Benson, Christian Requena-Mesa, Claire Robin, Lazaro Alonso, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Forecasting localized weather impacts on vegetation as seen from space with meteo-guided video prediction. *arXiv preprint arXiv:2303.16198*, 2023.

Sandra C Chapman, Nicholas W Watkins, and David A Stainforth. Warming trends in summer heatwaves. *Geophysical Research Letters*, 46(3):1634–1640, 2019.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

Codruț-Andrei Diaconu, Sudipan Saha, Stephan Günnemann, and Xiao Xiang Zhu. Understanding the role of weather data for earth surface forecasting using a convlstm-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1362–1371, 2022.

Yoan Fourcade, Aurélien G Besnard, and Jean Secondi. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, 27(2):245–256, 2018.

Chaonan Ji, Tonio Finke, Mélanie Weynants, Karin Mora, Vitus Benson, Fabian Gans, and Miguel Mahecha. Earth system minicubes towards revealing and predicting climate extremes. *in preparation*, 2024.

Gensuo Jia, Elena Shevliakova, Paulo Artaxo, De Noblet-Ducoudré, Richard Houghton, Joanna House, Kaoru Kitajima, Christopher Lennard, Alexander Popp, Andrey Sirin, et al. Land-climate interactions. 2019.

Klaus-Rudolf Kladny, Marco Milanta, Oto Mraz, Koen Hufkens, and Benjamin D Stocker. Enhanced prediction of vegetation responses to extreme drought using deep learning and earth observation data. *Ecological Informatics*, pp. 102474, 2024.

John F Kolen and Stefan C Kremer. *A field guide to dynamical recurrent networks*. John Wiley & Sons, 2001.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuchi Ma, Shuo Chen, Stefano Ermon, and David B Lobell. Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 301:113924, 2024.

Francesco Martinuzzi, Miguel D Mahecha, Gustau Camps-Valls, David Montero, Tristan Williams, and Karin Mora. Learning extreme vegetation response to climate forcing: A comparison of recurrent neural network architectures. *EGUsphere*, 2023:1–32, 2023.

Hanna Meyer and Edzer Pebesma. Predicting into unknown space? estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633, 2021.

Hanna Meyer and Edzer Pebesma. Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208, 2022.

Hanna Meyer, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, 411:108815, 2019.

J Eamonn Nash and Jonh V Sutcliffe. River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

David M Olson and Eric Dinerstein. The global 200: Priority ecoregions for global conservation. *Annals of the Missouri Botanical garden*, pp. 199–224, 2002.

Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, et al. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):225, 2020.

Pierre Ploton, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, Nicolas Bayol, et al. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature communications*, 11 (1):4540, 2020.

Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen, and Jukka Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Christian Requena-Mesa, Vitus Benson, Markus Reichstein, Jakob Runge, and Joachim Denzler. Earthnet2021: A large-scale dataset and challenge for earth surface forecasting as a guided video prediction task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1132–1142, 2021.

Claire Robin, Christian Requena-Mesa, Vitus Benson, Lazaro Alonso, Jeran Poehls, Nuno Carvalhais, and Markus Reichstein. Learning to forecast vegetation greenness at fine resolution over africa with convlstms. *arXiv preprint arXiv:2210.13648*, 2022.

Esther Rolf. Evaluation challenges for geospatial ml. *arXiv preprint arXiv:2303.18087*, 2023.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Roozbeh Valavi, Jane Elith, José J Lahoz-Monfort, and Gurutzeta Guillera-Arroita. blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Biorxiv*, pp. 357798, 2018.

A APPENDIX

Model training We employ a ConvLSTM architecture (Shi et al., 2015) following the original encoding-forecasting setup since ConvLSTM and LSTM architectures have been frequently used in vegetation forecasting tasks (Diaconu et al., 2022; Robin et al., 2022; Martinuzzi et al., 2023; Kladny et al., 2024). The model consists of two blocks for encoding and two blocks for the forecasting network, totaling 1 Mio. parameters. During the training, we employ a teacher forcing method (Kolen & Kremer, 2001), meaning randomly substituting some prior predictions with ground truth as inputs to the autoregressive model to expedite the convergence. We train our model on 10 epochs, using the AdamW optimizer (Loshchilov & Hutter, 2017), with a learning rate of 10^{-4} for the first epoch, then 10^{-5}, and finally 10^{-6} after the 7th epoch using a multi-step learning rate on a single A100 GPU. We use a masked MSE loss, removing the missing, cloudy or non-vegetation pixels.

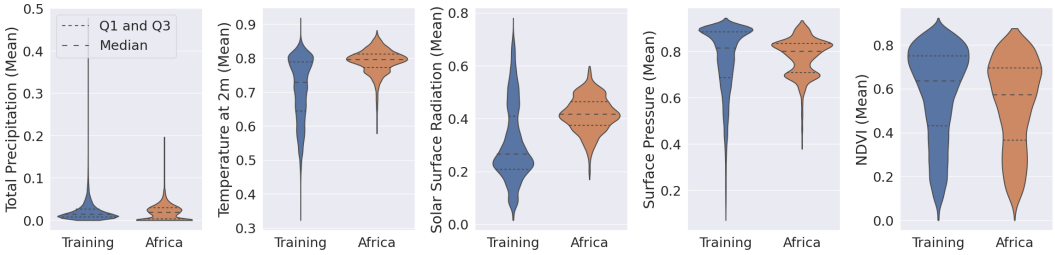

Figure 3: Distribution of several meteorological variables and NDVI after rescaling between the training set and the African test set of the *'w/o Africa'* splitting method.

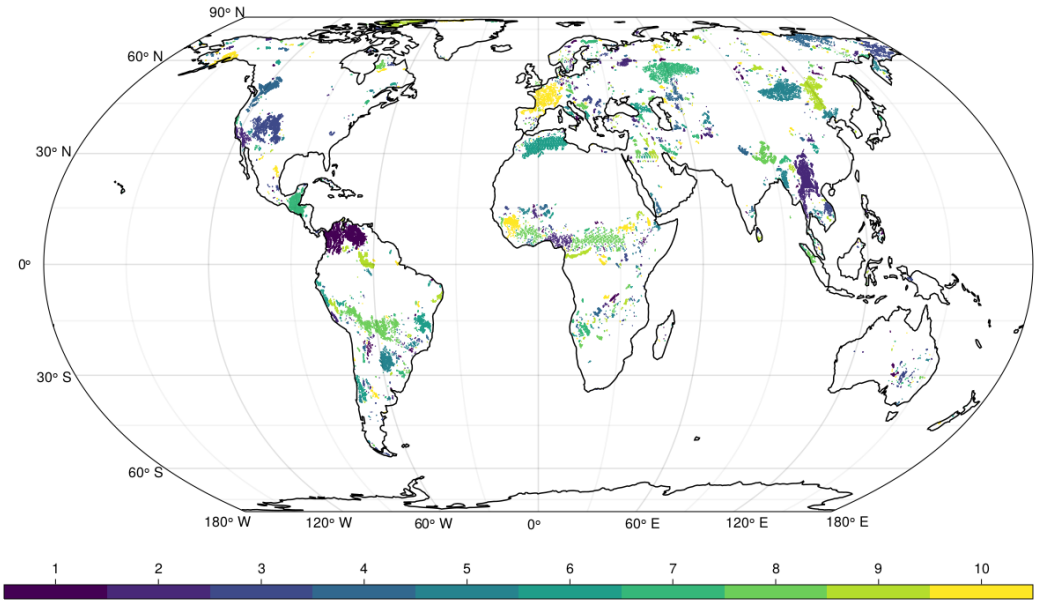

Figure 4: Map of the samples' location, with colors corresponding to the 10-fold split utilized to mitigate spatial auto-correlation.

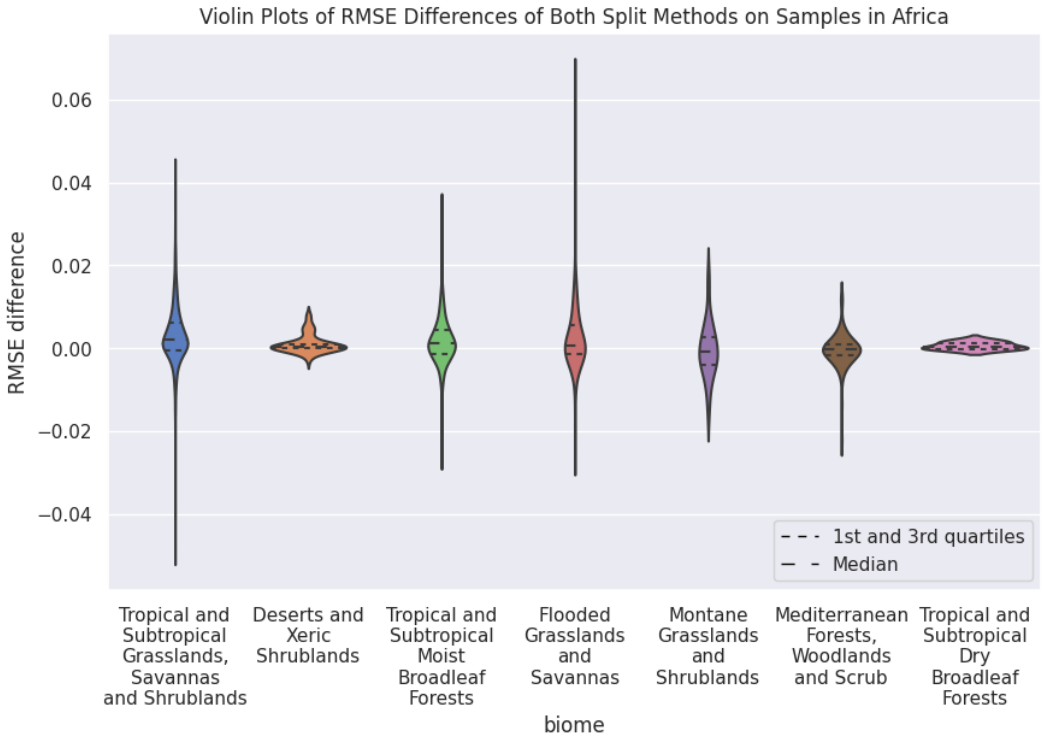

Figure 5: Violin Plots of RMSE Difference of Both Split Methods on African Samples. For each sample, we compare the difference of RMSE between both split methods. The violin plot shows that the error is Gaussian across the samples.