Project: **1444**
Project title: **EXPECT**
Principal investigator: **Étienne Plésiat**
Report period: **2025-05-01 to 2026-04-30**

Based on the work carried out in the previous project year on the infilling of the HadEX-CAM dataset using a U-Net architecture with partial convolutions[1], the approach has been extended to the reconstruction of a global precipitation dataset, namely the Global Precipitation Climatology Centre (GPCC) dataset[2]. The objective of this new phase was to assess whether the methodology previously developed for European climate extreme indices could be transferred to global monthly land precipitation fields, which present a larger proportion of missing values and a more complex spatial structure.

The GPCC Full Data Monthly Version 2022 dataset[3] provides monthly global land precipitation fields for the period 1891-2020. It is based on rain-gauge observations and is available on regular latitude–longitude grids at different spatial resolutions. As with many observational climate datasets, the spatial completeness and quality of the fields vary over time, especially during the earlier part of the record and in regions with sparse station coverage. It is therefore necessary to fill these data gaps using robust infilling techniques, such as the deep-learning methods developed at DKRZ.

The first stage of the work consisted of applying the same U-Net-based reconstruction strategy that had been successfully developed for HadEX-CAM. The model uses partial convolutions, allowing the convolutional filters to operate only on valid input values while ignoring missing or masked regions. This makes the architecture particularly suitable for climate datasets with large and heterogeneous regions of missing values. A new version of the U-Net model was trained, including circular padding to account for longitudinal boundary conditions on the sphere. The model architecture comprised 12 layers and 2048 channels in the bottleneck layer. Training was performed using eight historical models from the CMIP6 HighResMIP archive, and the model was evaluated against a test set of withheld HighResMIP data as well as against ERA5.

The evaluation, based on several metrics including the root mean square error (RMSE) and the Spearman rank-order correlation coefficient (SROCC), showed very good reconstruction skill. The U-Net model performed particularly well in comparison with a simpler baseline method, namely inverse distance weighting (IDW). These results indicate that the partial-convolution U-Net approach can be successfully extended to more complex reconstruction tasks involving global precipitation fields.

Nevertheless, the U-Net approach has structural limitations when applied to global climate fields. Its convolutional filters are formulated in Euclidean space and assume translation invariance, which is not strictly valid for data defined on the sphere. For global latitude–longitude grids, this can lead to inconsistencies and reconstruction artefacts, particularly in regions affected by the convergence of meridians. U-Nets also require fields on regular grid topologies, which limits their flexibility for representing global data on more natural spherical or irregular discretizations. Moreover, because convolutions are local operations, long-range spatial dependencies such as teleconnections or large-scale coherent precipitation patterns can only be captured indirectly, often requiring deeper and more computationally demanding architectures.

To overcome these limitations, a new Graph Neural Network (GNN) framework was implemented. Instead of treating the climate field as a conventional image, the method represents the data on a graph defined over spherical meshes. The regular latitude–longitude precipitation fields, together with positional information and missing-value masks, are encoded onto a hierarchy of multi-resolution icosahedral meshes. Connections between grid and mesh nodes are based on spatial overlap and include geometric information such as great-circle distance and relative direction. Information is then propagated across the mesh through message passing, allowing the model to learn spatial relationships while respecting the topology of the global domain. The reconstructed fields are finally projected back onto the target latitude–longitude grid. By operating on a spherical graph, this framework provides a more natural representation of global geometry and a more flexible treatment of both local and long-range spatial dependencies.

A substantial portion of the computational effort was devoted to model optimization. This involved an extensive search across a broad hyperparameter space, including learning rate, number of hidden channels, and network depth. Many distinct configurations were explored in order to identify models that yielded the best reconstruction performance while remaining computationally feasible. This optimization phase was essential because both U-Net and graph-based architectures are sensitive to the balance between model capacity, generalization ability, and training stability.

Model training was carried out using GPU nodes on Levante. The use of HPC resources was essential because the work involved global climate fields at 0.5º resolution, repeated experiments across different masking strategies, and extensive hyperparameter tuning. GPU acceleration made it possible to train deep-learning models at the required scale and to perform systematic optimization within a practical timeframe. CPU resources were mainly used for the dataloading and the preprocessing of gridded precipitation data.

The input data used for model development included simulations from eight distinct historical models within the HighResMIP archive. Validation was carried out using withheld random subsets from the same data source, allowing reconstruction accuracy to be assessed under controlled conditions where the true values were known. Further evaluation were carried out through cross-validation against the ERA5 reanalysis product. This additional test was designed to evaluate whether the trained models generalized beyond the specific data distribution used during training.

Overall, the GNN model performed slightly better than the U-Net, although the improvement was moderate (see Table below). However, the GNN approach is considerably more flexible by design and opens new opportunities for future infilling tasks, including the use of spatially irregular data sources such as station observations. It also provides a promising framework for other climate applications involving multivariate and spatiotemporal learning.

Evaluation	**Pr** (mm/month)	
Test set	**RMSE**	**SROCC**
IDW	42.50	0.95
U-Net	23.21	**0.98**
GNN	**22.71**	**0.98**

Evaluation	**Pr** (mm/month)	
ERA5	**RMSE**	**SROCC**
IDW	39.79	0.95
U-Net	25.97	**0.98**
GNN	**25.65**	**0.98**

The newly infilled GPCC datasets are currently being prepared for release to the research community, while the methodology and results are being consolidated for publication.

References

[1] Plésiat, É., Dunn, R.J.H., Donat, M.G. et al. Artificial intelligence reveals past climate extremes by reconstructing historical records. Nat Commun 15, 9191 (2024).

[2] https://www.dwd.de/EN/ourservices/gpcc/gpcc.html

[3] Schneider, U. et al., DOI: 10.5676/DWD_GPCC/FD_M_V2022_100 (2022)