https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

From climate projections to trajectory-conditioned local predictions

Simon L. L. Michel[1*], Kristian Strommen[1], Hannah M. Christensen[1]
[1]Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Parks Road, Oxford, OX1 3PU, United Kingdom
Correspondence to: Simon L. L. Michel (simon.michel@physics.ox.ac.uk)

Abstract. Regional climate projections are usually sharpened by constraining models' forced responses toward observations, a step that filters out internal variability and works well mainly where the signal-to-noise ratio is high. Here we side-step this limitation with deep learning, treating internal variability not as noise to remove but as part of the future to predict and of the uncertainty to report. A U-Net trained on 290 simulations from 40 CMIP6 Earth System Models maps the 1980 to 2025 evolution of surface air temperature onto decadal-mean monthly climatologies from the 2030s to the 2090s under SSP2-4.5, keeping the forced response and internal variability together. Tested on entirely withheld model families, the forecasts are skilful and empirically calibrated out-of-sample. Within a model's own ensemble, a forecast predicts the future of the specific member it was given more accurately than the futures of that model's other members, a skill that persists to the 2090s, notably over the subpolar North Atlantic, Europe, North America and Southern Ocean. The forecasts also generalise to a next-generation eddy-rich model whose resolved physics lies outside the training distribution, supporting application to real-world data. When applied to observations, the framework reduces uncertainty in global-mean warming similarly to observational constraints, while also providing spatially coherent local predictions. This shows that our deep learning approach can estimate the distribution of future local climate, conditioned on the recent observed trajectory, encompassing uncertainty arising from both forced response and internal variability realisations. This extends observationally constrained projection toward calibrated predictions at the local scale and seasonal resolution where adaptation decisions are made.

1 Introduction

Twenty-first century climate projections remain substantially uncertain at the regional and local scales relevant for adaptation (Deser et al. 2020, Pörtner et al. 2022). For many temperature-related metrics this uncertainty is larger in CMIP6 than in CMIP5 (Lehner et al. 2020), driven by divergent representations of cloud feedback (Zelinka et al. 2020, 2022), cloud-aerosol interactions (Meehl et al. 2020), and the sea-surface temperature pattern effect on these and other climate feedbacks (Dong et al. 2020).

The main response to this spread has been observational constraints (OCs), which link present-day observables to the Earth System Model (ESM) forced response, for example by regression on large-scale or regional quantities (e.g., Cox et al. 2018, O'Reilly et al. 2024, Portmann et al. 2025, 2026) or by Bayesian inference from global to local scales (O'Reilly et al. 2024, Qasmi and Ribes 2022, Ribes et al. 2021, 2022, 2025). In each case the prior is built at the ESM level: one forced response per ESM, estimated after filtering internal variability.

However, the quantity society experiences, and that adaptation must plan for, is a single realised trajectory (Sippel et al. 2019). The forced response approximates it well only where it dominates internal variability (i.e., high signal-to-noise ratio, S/N), such as for global mean surface temperature (GMST) and several tropical regions (Fig. 1a,b). Where S/N is low, such as in several mid-latitude locations (King and Harrington 2018), the low-frequency internal-variability spread around the forced response is large (Fig. 1c–e), so the uncertainty that matters operationally is much wider than a constrained forced response implies. Even a perfectly estimated forced response is then not what a location will experience, and the forced response of the real world cannot itself be estimated: a single realisation exists, with no ensemble to average over. For quantities with strong internal variability we therefore do not know whether the observed record has run close to the forced response or far from it. Low-frequency internal variability is not just noise to remove but a structural part of any multi-decadal trajectory (Deser et al. 2020), and it is exactly what adaptation has to plan for. Adding it back after constraining, as a noise term on the forced response (Ribes et al. 2021, Qasmi and Ribes 2022), restores variance but not the specific low-frequency state a region or location will actually follow.

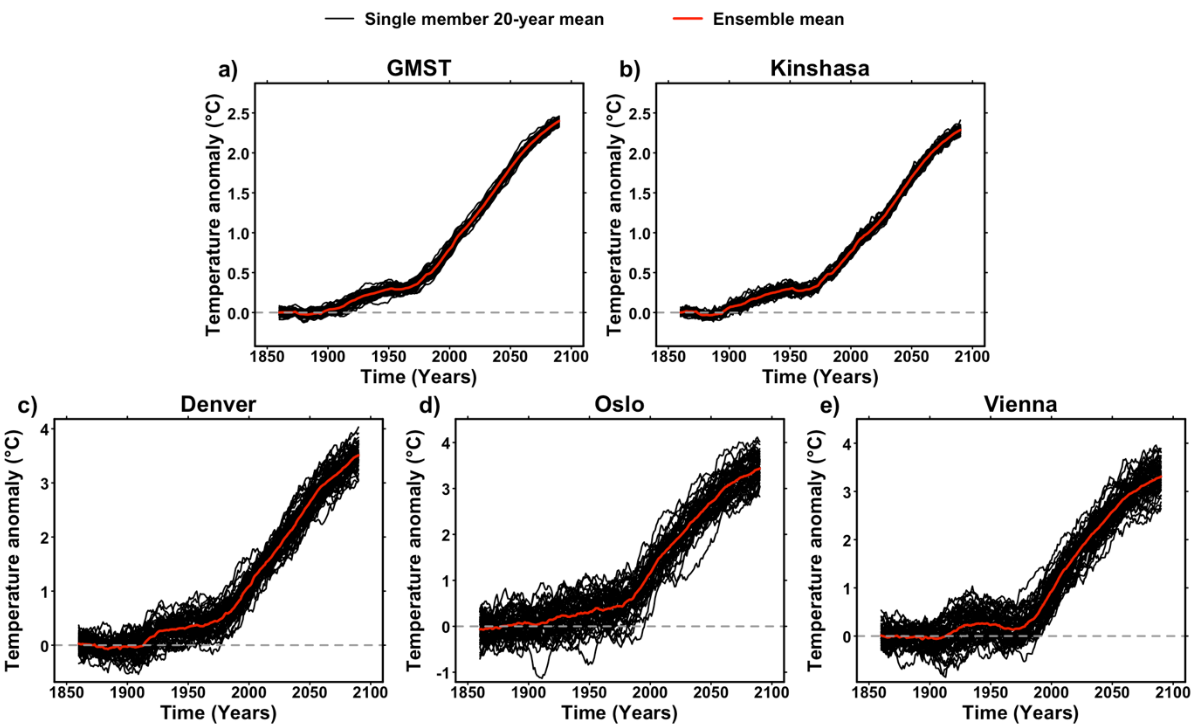


Figure 1: Examples of global and local forced responses. The plot shows 20-year cantered mean temperature averages for 50 members of merged historical and SSP2-4.5 simulations from the MPI-ESM1-2-LR Earth System Model (ESM). Data is represented for global mean surface temperature (GMST) (a); a city with a high signal-to-noise temperature regime: Kinshasa, Democratic Republic of the Congo (b); and three cities with low signal-to-noise temperature regimes: Denver, USA (c), Oslo, Norway (d), and

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Vienna, Austria (e). Black lines are single member 20-year temperature averages, and red lines are ensemble member mean of these 20-year temperature averages.

This creates an additional data-related problem for any prior built on one response per ESM. Using a single member risks locking in one arbitrary trajectory (Fig. 1c–e, Sippel et al. 2019). Estimating the forced response from a large ensemble avoids this, but only a handful of CMIP6 ESMs provide large ensembles (Maher et al. 2021), so restricting the prior to them leaves too few models to build a reliable constraint. This runs into a known weakness: with few models and many candidate predictors, apparent constraints can arise by chance or through shared biases and then fail on unseen models (Caldwell et al. 2014, Sanderson et al. 2021, Nowack and Watson-Parris 2025). Taking one ensemble-mean per ESM across the full archive does not resolve this either, because ensemble sizes are uneven: it mixes a few well-estimated forced responses, from the large-ensemble ESMs, with many noisy single-member draws. Most constraint- and prediction-based methods work from one forced-response per ESM (O'Reilly et al. 2024, Qasmi and Ribes 2022) and inherit this low-S/N problem, which is worse for regional precipitation (O'Reilly et al. 2024) and hydrological extremes (Bevacqua et al. 2023). Adding internal variability back after constraining, as a noise term on the forced response (Ribes et al. 2021, Qasmi and Ribes 2022), restores variance but not the specific low-frequency state a region will actually follow.

Internal variability is also tied to the forced response itself: the size of local temperature variability changes as the climate warms, in a pattern that differs by region and by model, growing over tropical land and shrinking at high latitudes (Screen 2014, Olonscheck et al. 2021). Estimating the local mean response and its spread across ESMs together is therefore a necessary step to estimate correctly how local variability itself changes, which reducing everything to a noise-free forced response discards by design.

Here, we forecast trajectory-specific decadal-mean monthly climatologies of local surface air temperature conditioned on the recent observed trajectory. In this way, the forced response and low-frequency internal variability remain embedded together at each location. To achieve this, we train a U-Net combined with dilated temporal convolutions (TC-UNet) on 290 historical+SSP2-4.5 simulations from 40 CMIP6 ESMs (Supplementary Table 1), mapping the 1980–2025 spatio-temporal evolution of monthly surface-air-temperature anomalies onto their future decadal-mean monthly climatologies from the 2030s to the 2090s. Every realisation is a training example, so the network sees both forced responses and a diversity of internal-variability realisations consistent with each ESM's physics. Skill is assessed with leave-one-model-family-out cross-validation, that is, on unseen physics rather than in-sample fit (Nowack and Watson-Parris 2025), and forecast uncertainty is built empirically from leave-one-model-family-out cross-validation errors, so the posterior spread carries both inter-ESM differences and internal variability.

Our TC-UNet generalises across ESM families, outperforms ensemble-aggregation baselines, provides skill for each specific trajectory in the largest ESM ensembles, and extends to unseen CMIP7-like HadGEM3-GC5 simulations, including a higher-ocean-resolution configuration (10km) coupled with a high-resolution atmosphere (<10kn) whose resolved physics lies outside the training distribution. Applied to reanalysis-derived inputs, it yields empirically calibrated uncertainty ranges whose GMST reduction is comparable to recent Bayesian constraints (Ribes et al. 2021), while delivering local estimates that, by

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

construction, account for internal-variability realisations in both prior and posterior. This extends the prediction framing of Li et al. (2025) from the global mean to the societally relevant local scale.

2 Data and Methods

2.1 Datasets

2.1.1 CMIP6 data

We use CMIP6 simulations (Eyring et al. 2016) combining the historical experiment (1850–2014, observed forcings) with SSP2-4.5 (2015–2100; O'Neill et al. 2016). For each ESM, historical and SSP members with matching identifiers (e.g., r1i1p1f1) are concatenated into continuous monthly trajectories, yielding 290 simulations from 40 ESMs (Supplementary Table 1), each covering 1850–2100 initially. The analysis is restricted to SSP2-4.5 because the approach is data-hungry: this scenario offers the largest pool of ESMs and members and is the most consistent with current emission trajectories and commitments (Hausfather and Peters 2020, Hausfather 2025, UNEP 2025, Venmans and Carr 2024). All monthly surface-air-temperature (tas) fields are bilinearly interpolated to a common 2°×1.875° grid.

2.1.2 Observational and reanalysis data

To apply the learned historical-to-future mapping to real-world data, we use one reanalysis and one observational product. The ERA5 reanalysis (Hersbach et al. 2020) is the most physically consistent analogue of the training data, providing 2-meter air temperature over both land and ocean, but it begins in the late twentieth century and offers no preindustrial reference. We additionally apply the framework to BEST (Rohde and Hausfather 2020), consistent with the IPCC GMST definition (Gulev et al. 2021) and included for comparability with constraint studies (Ribes et al. 2021). BEST is however physically inconsistent with the training data: over ocean it contains sea surface temperatures, which warm more slowly and vary differently from the near-surface air temperature the network learnt. BEST forecasts are therefore reported as a sensitivity test rather than the primary application. Blending air and ocean temperatures during training is not a viable alternative: the network learns full spatial maps, so the diverging land-ocean masks across ESMs would contaminate training, unlike grid-cell-level methods that can use blended fields directly (Qasmi and Ribes 2022).
For both products, monthly fields over 1980–2025 are regridded to the common 2°×1.875° grid and expressed as anomalies relative to the 1995–2014 monthly climatology.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

2.1.3 Anomaly space

All fields are expressed as anomalies relative to a seasonally varying 1995–2014 climatology, computed per dataset and per calendar month. A preindustrial baseline (e.g. 1850–1900), standard in IPCC assessments (e.g., Seneviratne et al. 2021), is infeasible in a spatially resolved real-world setting: the global-mean offset from preindustrial is well constrained, but the spatial structure of the preindustrial climatology is not observable, highly uncertain when inferred, and beyond the reach of modern reanalyses. The 1995–2014 reference is instead fully covered by CMIP6, ERA5, and BEST, ensuring consistent treatment across ESMs, reanalyses, and observations.

This choice retains the information that drives predictive skill: recent forced trends, spatial warming patterns, and low-frequency internal variability all survive in the anomaly fields. Learning in absolute-temperature space would instead let the network exploit ESM-specific mean states, artificially easing the task and yielding unrealistically narrow uncertainty (observed in preliminary experiments, not shown).

When GMST is reported relative to preindustrial (e.g. Seneviratne et al. 2021, Gulev et al. 2021), the published 0.85 °C offset of 1995–2014 relative to 1850–1900 (IPCC 2021, Lee et al. 2021) is applied a posteriori, separately from the anomaly framework used for training.

2.2 Forecasting future decadal-mean climatologies

2.2.1 Individual prediction

For each decade d from the 2030s (2030–2039) to the 2090s (2090–2099), we learn a data-driven mapping of the recent evolution of the climate to the future decadal-mean monthly climatology, at the level of individual ensemble members. The network learns to utilise information from both the ESM's forced trajectory and the member's particular realisation of internal variability to predict future climate. The prediction therefore encodes both the ESM's forced trajectory and the member's particular likely realisation of internal variability over the future decades.

For simulation m, the input, X_m, is the 46-year sequence of maps of monthly surface temperatures over 1980–2025:

$$X_m = \{T_m(t,s): t \in [1980,2025]\}, \tag{1}$$

Where $T_m(t,s)$ is the local temperature at a monthly timestep t and grid cell location s for simulation m. The target, $Y_m^{(d)}$, for decade d is the map of the future monthly climatology:

$$Y_m^{(d)} = \overline{T_m^{(d)}}(s), \tag{2}$$

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

where the overbar denotes a calendar-month-wise average over decade d. We train a decade specific predictor:

$$\hat{Y}_m^{(d)} = F^{(d)}(X_m). \tag{3}$$

Ensemble sizes vary across ESMs, but the task stays well-posed because each simulation is treated as an independent climate trajectory and generalisation is assessed at the ESM family level (see below). Because the input spans 46 years, $F^{(d)}$ can exploit persistent low-frequency modes (e.g., Atlantic Multidecadal Variability, Yeager et al. 2021, Zhang et al. 2019, or Pacific Decadal Variability, Power et al. 2021) carried in the member trajectory, which keep driving short- to long-term internal climate variability (Deser et al. 2020, Deser 2020).

2.2.2 Seed averaging

Training is subject to stochasticity from weight initialisation, batch ordering, and early stopping, which is unrelated to physical uncertainty and occasionally yields offset predictions (observed at the regional scale for less than 5% of cases, not shown). We suppress it by training five independent seeds and using their mean as final predictions. All forecast errors, and uncertainty estimates below use this five-member mean.

To address skill under significant ESM differences, we introduce leave-one-model-family-out cross-validation (LOMFO-CV, Supplementary Table 1). The purpose of the LOMFO-CV is to remove the effect of too close proximities of models developed within the same modelling centre, which share substantial parametrisation schemes, coupling strategies, and components of the dynamical core (Brunner et al. 2020, Brunner and Sippel 2023, Knutti et al. 2017, Michel et al. in rev.). Let G_k be all simulations from family k. For family k,

$$D_{train}^{(k)} = \cup_{j \neq k} G_j, D_{test}^{(k)} = G_k. \tag{4}$$

i.e. we train on all simulations from all other model families and test performance on the held-out family k. This is repeated for all model families (Supplementary Table 1). For any derived quantity from the forecast (e.g., GMST, seasonal mean, regional average, local annual average), errors are evaluated only on held-out families,

$$e_m^{(d)} = \hat{Y}_m^{(d)}(s) - Y_m^{(d)}. \tag{5}$$

and pooled across LOMFO-CV folds to obtain the empirical distribution from raw errors:

$$E^{(d)}(s) = \{e_m^{(d)}(s): m = 1, \dots, M\}. \tag{6}$$

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Pooling members with equal weight would over-represent large-ensemble families (e.g. 50 members of MPI-ESM1-2-LR against a single-member ESM such as CMCC-CM2-SR5, Supplementary Table 1). Hence, unweighted pooling biases the empirical quantiles toward the internal-variability spread of the largest ensembles. This exchangeability problem was also noted for large ensemble-aware ESM weighting (Merrifield et al. 2020) and emergent-constraint sampling (Tokarska et al. 2020, Brunner et al. 2020).

We therefore build the error distribution by ESM-level Monte Carlo resampling. For each fold k and decade d we draw one error $e_m^{(d)}(s)$ uniformly from each contributing ESM, giving a balanced 40-member sample (*i.e.*, one member for each of the ESMs). We then compute 5th and 95th percentiles and repeat B=10,000 times. The reported bounds are the mean of resampled quantile pairs,

$$[\hat{Y}_{obs}^{(d)}(s) + \overline{Q_{0.05}(E_b^{(d)}(s))}\,, \hat{Y}_{obs}^{(d)}(s) + \overline{Q_{0.05}(E_b^{(d)}(s)\,)}], \tag{7}$$

where $E_b^{(d)}$, is the b-th balanced resample and Q_α the α-quantile. This weights each ESM equally as a unit of structural uncertainty while preserving within-family spread as the carrier of internal variability, so the interval reflects the diversity of CMIP6 physics rather than which centres happened to run large ensembles. The approach is still imperfect as single-member ESMs see their realisation sampled across all Monte-Carlo draws. This limitation, tied to data availability, and possible overcoming strategies, will be highlighted in the discussion section of the paper.

Because each fold tests against a different family and a specific internal-variability realisation, the resulting error distribution carries both irreducible internal variability and ESM-to-ESM forecast error. The five-seed averaging (Section 2.2.2) filters-out optimisation noise, not these physical sources.

2.3 Neural Network

Each decade-specific predictor $F^{(d)}$ is a two-stage neural network comprising a temporal feature extractor (Fig. 2a), and a U-Net (Fig. 2b, Ronneberger et al. 2015) and is referred to as TC-UNet (combined temporal convolutions and U-Net) in the following (Fig. 2). It maps the 46-year input field X to the predicted monthly climatology $\hat{Y}_{obs}^{(d)}$. The same architecture is used for the cross-validation described in section 2.2.3, and for the final real-world application, where all simulations are used in training. The following describes the logic and general steps of the architecture. Additional architectural and parameter details are given in Fig. 2 and in Appendix A.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

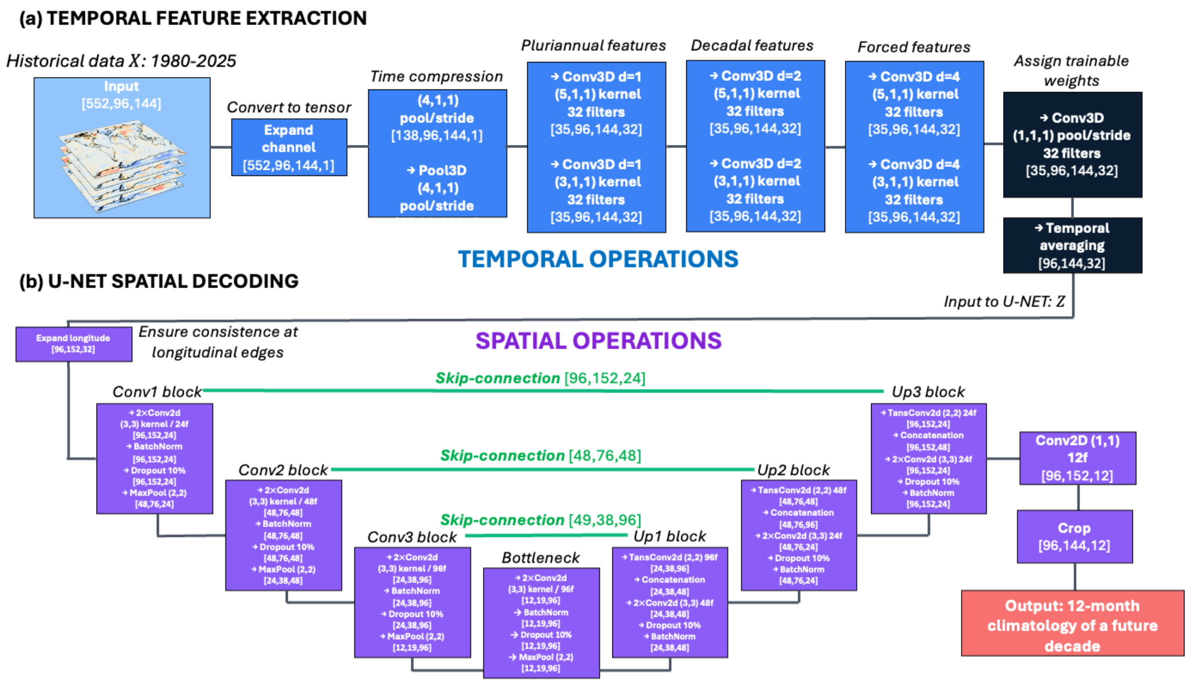


Figure 2: Sketch of the TC-UNet neural network architecture. The input is a full spatio temporal field of monthly data over 1980–2025. The blue light path indicates the temporal feature extraction phase (a). Dark blue indicates the intermediate phase with temporal compression. The purple path is the U-Net part, which encodes-decodes the temporal feature space (b) into a 12-month future climatology for a given target decade (red). The U-Net skip-connections are represented in green. While providing a schematic representation of the architecture, the figure also provides details on trainable weight and architectural workflow (inside rectangles).

2.3.1 Temporal feature extraction

At each grid cell (i, j) the monthly sequence $x_{(i,j)} \in \mathbb{R}^T$, $T = 552$, is first reduced by temporal average pooling to emphasise interannual-to-decadal variability. The pooled sequence passes through a stack of dilated 1-D temporal convolutions, implemented as 3-D convolutions with kernel $(k, 1,1)$ so that weights act only along time and are shared across all grid cells (Fig. 2a). Dilation factors of 1, 2, and 4 let the network capture multiannual, decadal, and trend-like variability without long kernels or many parameters. The encoder outputs a latent representation $Z = \mathbb{R}^{C\times H\times W}$ with $C = 32$ channels, summarising the global temporal evolution most predictive of $Y^{(d)}$.

2.3.2 Spatial decoding

Z is passed to a 2-D U-Net encoder-decoder that reconstructs the seasonally resolved climatology for decade d. Down- and up-sampling integrate information across spatial scales while skip connections preserve fine scale details from the encoder branch (Fig. 2b). The outputs are the 12 monthly climatological fields $\hat{Y}^{(d)} \in \mathbb{R}^{12\times H\times W}$. Separating temporal extraction from

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

spatial reconstruction lets the TC-UNet map recent variability, forced and internal, directly into coherent future climatological patterns.

2.3.2 Training

Each $F^{(d)}$ has $\sim 3.76 \times 10^6$ trainable parameters and is trained with the Adam optimiser (Kingma and Ba 2015) at an initial learning rate of 10^{-3}, batch size 4, minimising the area weights mean absolute error between $\hat{Y}^{(d)}$ and $Y^{(d)}$ over all grid points and month. Two factors guard against overfitting. First, skill is evaluated entirely out-of-sample under LOMFO-CV, so a network that fit training-member noise rather than transferable structure would generalise poorly to held-out families. Second, reducing the decoder to two down- and up-sampling levels systematically degraded held-out MAE across a representative subset of folds (not shown), indicating that the chosen capacity is not excessive for the task.

For all neural networks, data is systematically randomly split into 80% of training and 20% of validation. Overfitting to the specific training split is limited by using a plateau-based learning rate scheduler and early stopping, both monitored on the validation loss. Beyond regridding, no detrending or preprocessing is applied, so the network learns the direct mapping $X \longrightarrow Y^{(d)}$ from raw recent anomalies with respect to 1995–2014.

3 Results

3.1 Cross-validation skill across CMIP6 ESM families

Forecasts under SSP2-4.5 were produced with the LOMFO-CV framework of Section 2.2.3 across 25 ESM families (Supplementary Table 1). To test whether TC-UNet (Section 2.3) captures regional and seasonal temperature structure beyond conventional ensemble aggregation, we compare it against standard baselines under the identical protocol, so that every method is determined only on the training families and evaluated on the withheld one. The baselines are the ensemble mean (EM), the weighted ensemble mean (wEM), the analogue method (ANA), and the averages of the five and ten closest training members in plain (B5, B10) and weighted (wB5, wB10) forms, with proximity and weights defined by member-level distance in absolute GMST over 1980–2025, where each member across the LOMFO-CV folds is sequentially taken as the pseudo-observational product.

3.1.1 Global forecast errors

We computed area-weighted mean absolute errors (MAE) from the out-of-sample predictions, per grid cell and month, then aggregated over the annual cycle. Fig. 3 summarises these errors across the seven forecast decades. Including past trajectory

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

information through TC-UNet gives the lowest mean error of any method at every target decade, and the reduction relative to the strongest baseline remains stable with lead time, indicating that recent temperature evolution stays predictive of future decadal climatologies even at multi-decadal range. TC-UNet also attains lower MAE than the baselines for nearly all model families (22/23), so the learned relationships transfer well across distinct ESM formulations. Through the short (2030s–2040s) and medium (2050s–2070s) range its error distribution is the tightest of any method, broadening only slightly toward the end of the century: skilful performance is consistent across families rather than dominated by a few well-predicted ESMs, indicating robustness.

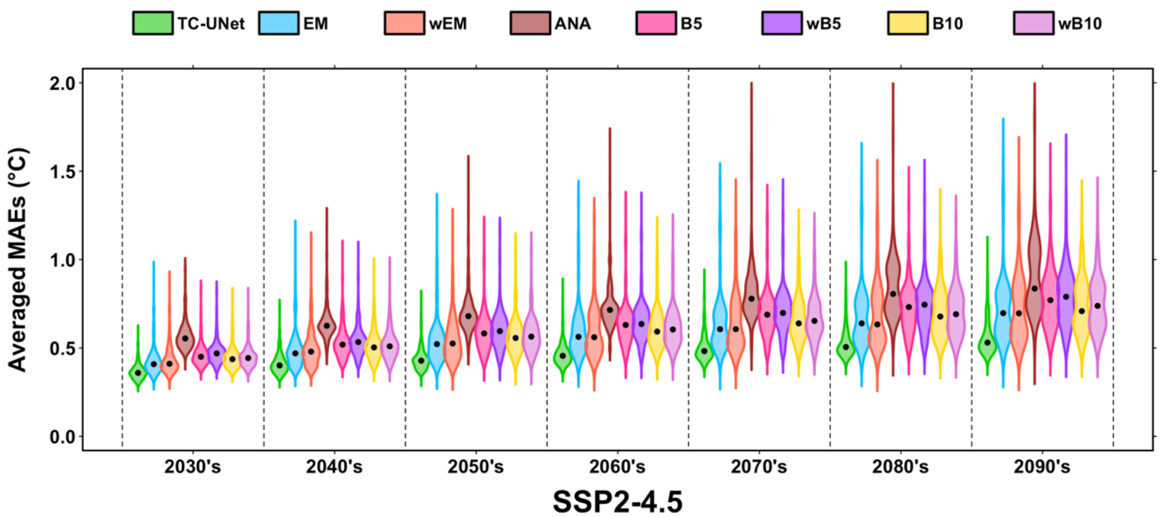
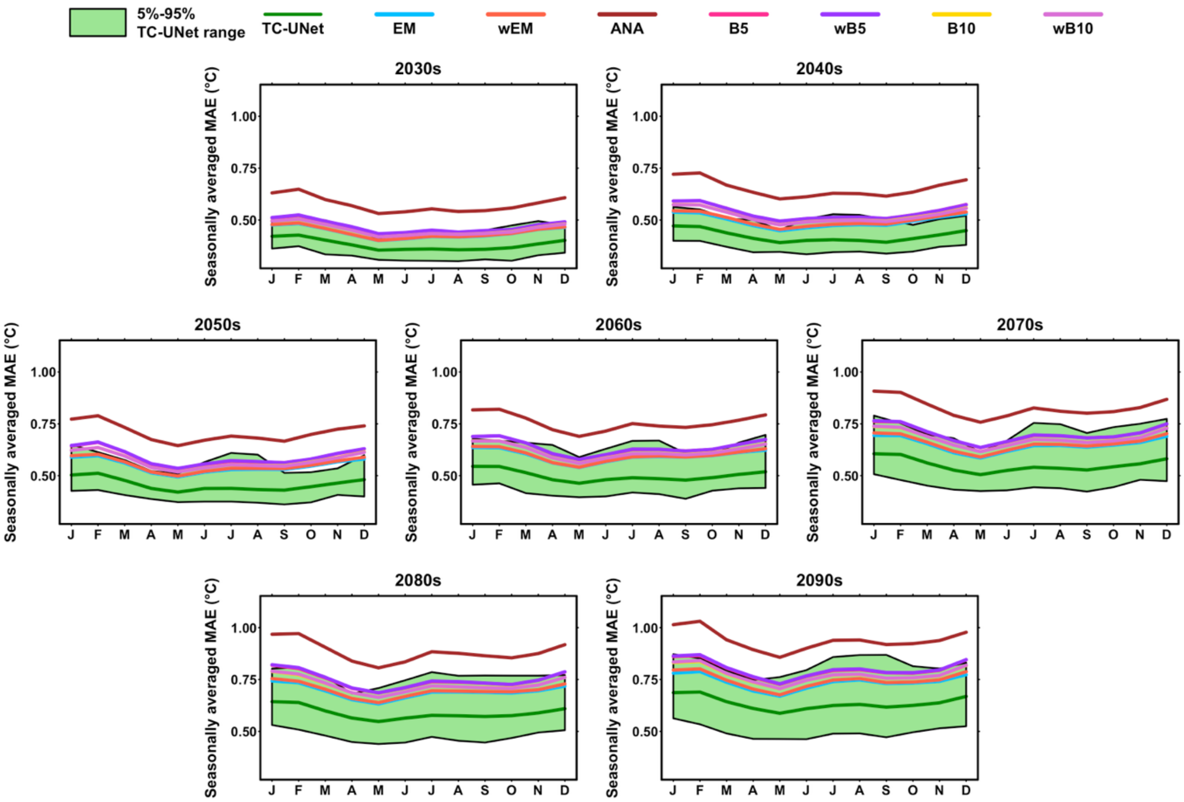


Figure 3: Global performance of TC-UNet and baseline ensemble methods. Spatially and annually averaged mean absolute errors (MAE) are represented for the 7 forecast decades and separated by vertical dashed lines: 2030s to the 2090s. Violins give MAEs for our TC-UNet (green) compared to: ensemble mean (EM, blue), weighted ensemble mean (wEM, red), analogue (i.e., best member, ANA, brown), mean of best 5 (B5, pink) and 10 (B10, yellow) training members and their weighted mean versions (wB5, purple; wB10, orchid). Baseline methods are computed from training data, and whether they are "best", as well as weights, are determined from their distance in absolute global mean surface temperature (GMST) with respect to the member to predict.

3.1.2 Monthly mean forecast errors

Because TC-UNet reconstructs full monthly climatologies, LOMFO-CV also resolves performance across the calendar year. The reduction in global MAE holds for every calendar month and forecast decade (Fig. 4). Errors rise modestly in boreal winter, as for the baselines, but the seasonal modulation of the TC-UNet error is markedly flatter than that of the analogue and ensemble methods. This indicates that the network captures not only the mean warming signal but part of its seasonal modulation, possibly including identified forced changes in the annual cycle (e.g., Cassou and Cattiaux 2016 for Europe).

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026



250 **Figure 4: Monthly distribution of mean absolute error (MAE). Monthly distribution of spatially averaged MAE across calendar months for each forecast decade. The TC-UNet prediction is indicated with the dark green line, whereas the 5%–95% distribution of all 290 individual LOMFO-CV errors is indicated by the green area. Solid lines give MAEs for our TC-UNet (green) compared to: ensemble mean (EM, blue), weighted ensemble mean (wEM, red), analogue (i.e., best member, ANA, brown), mean of best 5 (B5, pink) and 10 (B10, yellow) training members and their weighted mean versions (wB5, purple; wB10, orchid). For visual clarity, the 255 5%–95% intervals are only shown as the green-shaded areas for TC-UNet and are determined from leave-one-model-family-out cross-validation (LOMFO-CV). Baseline methods are computed from training data, and whether they are "best", as well as weights, are determined from their distance in absolute global mean surface temperature (GMST) with respect to the member to predict.**

3.1.3 Spatial distribution of annual mean forecast errors

We next computed annual-mean MAE at each grid cell. Fig. 5 shows these errors for short- (2030s–2040s), medium- (2050s–260 2070s), and long-range (2080s–2090s) forecasts, alongside the strongest baseline (EM). TC-UNet has lower error than EM across the large majority of land and ocean regions at all three ranges, with the relative improvement covering most of the globe including the tropics and southern oceans (Fig. 5, fourth column). The improvement is smallest, and locally non-significant, over Northern Hemisphere mid-latitude and polar regions, consistent with the low predictability imposed by strong internal variability there at both short (Hawkins and Sutton 2009) and long (Deser et al. 2020) ranges.

265 Together, these results appears to show that conditioning on past trajectories of internal variability, as in TC-UNet, improves the reconstruction of both large-scale and regional temperature structure beyond ensemble averaging or ESM-selection

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

approaches, even out to multi-decadal lead times (Figs. 3–5). In the following section, we perform several tests to clarify if TC-UNet is indeed using the information from past trajectories, rather than some other ESM-specific information.

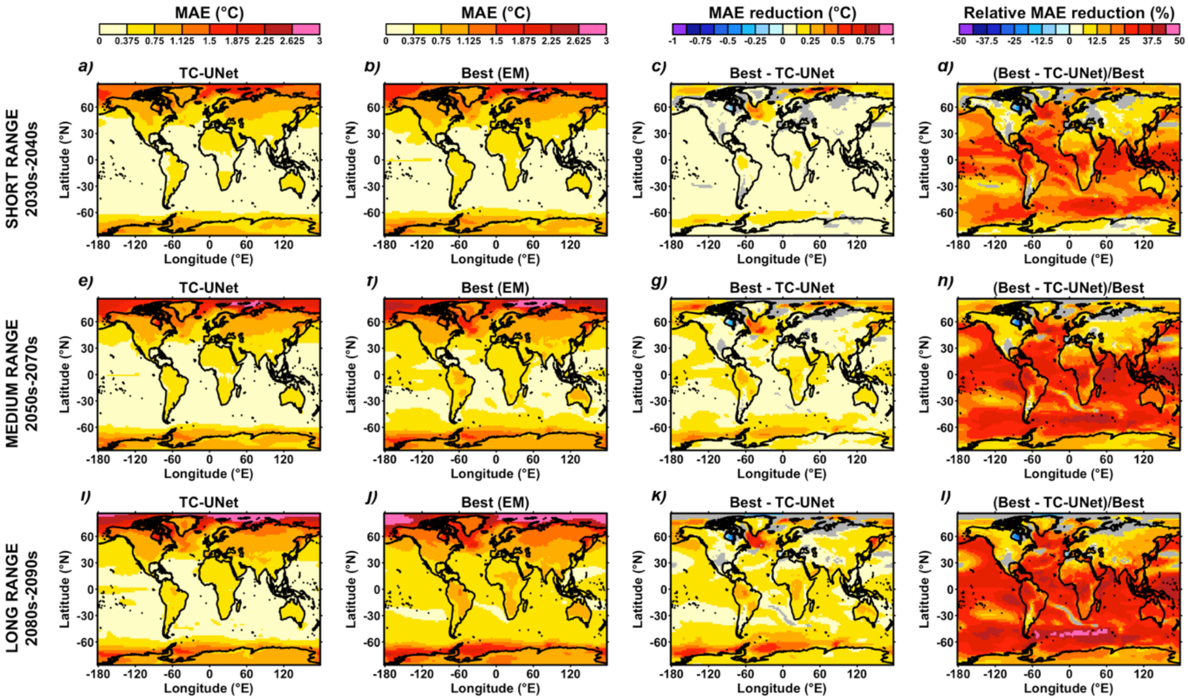


Figure 5: Spatial distribution of mean absolute error (MAE) averaged across calendar months. The MAE spatial distribution is average across forecast decades for three different groups: short range (2030s–2040s, a–d), medium range (2050s–2070s, e–h), long range (2080s–2090s, i–l). The first column gives the MAE distribution for TC-UNet for each time range group (a,e,i). The second column gives the MAE distribution for the best of the baseline methods (ensemble mean, EM) in terms of spatially averaged global MAE (Fig. 3) for each time range group (b,f,j). The third column gives the distribution of MAE difference between EM and TC-UNet for each time range group (a,e,i). The fourth column gives the distribution of relative MAE difference (in %) between EM and TC-UNet for each time range group (a,e,i).

3.2 Realisation-level analysis

Because each realisation is treated as a single sample in TC-UNet, it can exploit two kinds of information: ESM-specific information, translating an ESM's historical temperature patterns into its future states given its particular climate sensitivity and warming patterns; and trajectory-specific information, i.e. the signal, if any, that the particular 1980–2025 trajectory carries about the future for a given ESM. To separate the two, we retain the 12 ESMs with more than five members (Supplementary Table 1) and test whether the prediction for a given member outperforms predictions conditioned on the other members of the same ESM. Skill at the realisation level would show that the network exploits the state of low-frequency modes to refine predictions beyond the ESM-specific patterns it necessarily fits across realisations. No skill at this level would imply that the network exploits only the ESM-specific historical warming pattern, behaving as an OC method dressed in a deep learning prediction framework.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

3.2.1 Prediction skill beyond ESM-specific patterns

We compare two member-level errors under LOMFO-CV, for each of the 12 ESMs with more than five members. The self-prediction MAE compares each member's forecast to that same member's true future climatology, and averages over the ESM's members. The cross-prediction MAE compares each member's forecast to the true future climatologies of all the other members of the same ESM, and averages over all such forecast-truth pairs. If forecasts contained only ESM-level information, a forecast would be no closer to its own member's future than to any other member's, and the two errors would coincide. We quantify their difference by the relative MAE reduction: the cross-prediction MAE minus the self-prediction MAE as a percentage of the cross-prediction MAE, computed from area- and monthly-averaged errors.

The relative MAE reduction is positive for every of the 12 ESMs with more than five members and every forecast decade (Fig. 6). On average, TC-UNet forecasts are systematically closer to the future of the specific member whose historical trajectory was the input than to the futures of other members of the same ESM: the recent trajectory carries predictive information beyond the ESM-specific warming pattern. When globally averaged, the mean reduction is generally modest. It is largest at short range for most ESMs, typically 3 to 7% in the 2030s, and decays to 1 to 3% by the end of the century as the memory of the initial state fades and the forced signal increasingly dominates the target. However, GISS-E2-1-G diverges from this pattern, with reductions of 6 to 8% that persist and even grow through the 2090s. This diverging behaviour is expected given the composition of its CMIP6 ensemble: unlike the other ESMs, whose members differ only in initial conditions, the GISS-E2-1-G members span distinct physics versions with prescribed or prognostic atmospheric composition and in turn different forcings (Miller et al. 2021). This in turn shows that the similar pattern observed in other ensembles with fixed physics is effectively resulting from the different realisations of internal variability. Excluding GISS-E2-1-G, part of the TC-UNet skill is thus realisation-specific: the network does more than an observational constraint reformulated as deep learning as it does not only identify future patterns based on ESM-only specificities.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

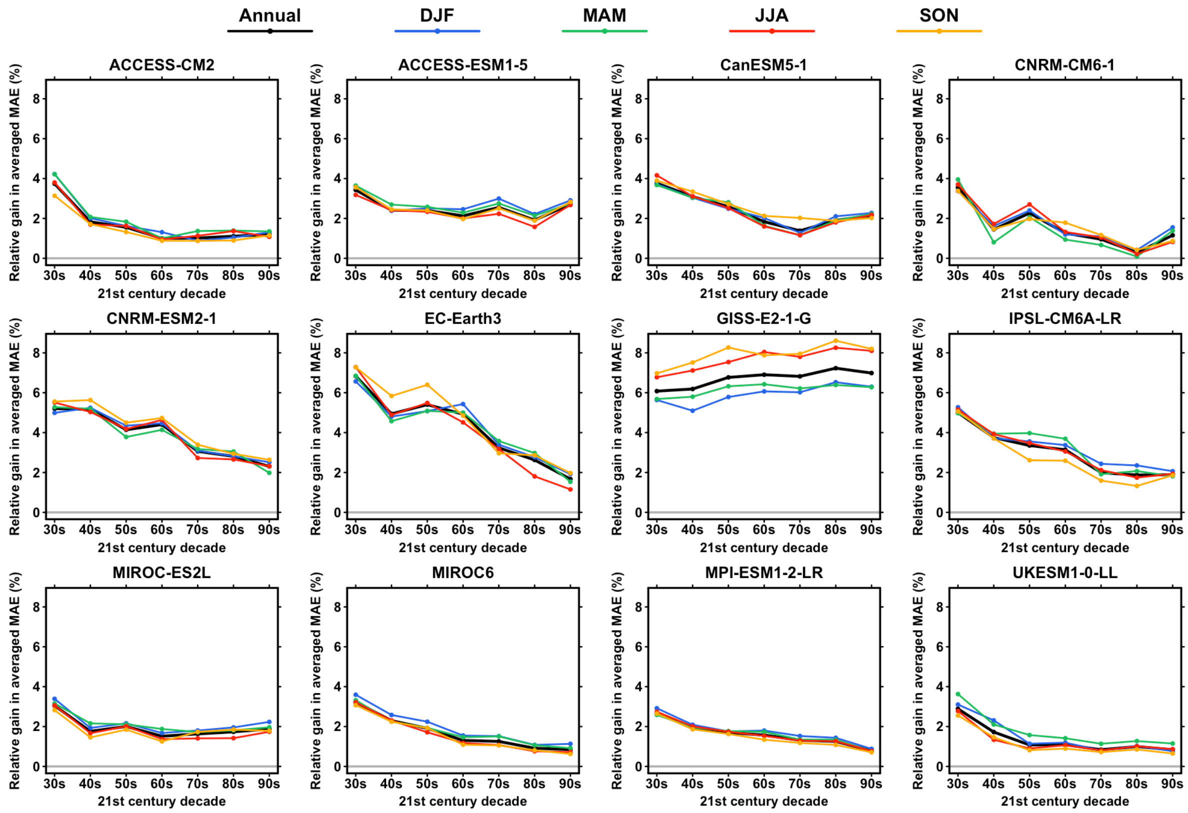


Figure 6: Relative Mean Absolute Error (MAE) reduction from conditioning on a specific realisation of internal variability. Each panel shows, for one ESM, the relative MAE reduction in area-weighted mean absolute error (MAE), defined as the mean cross-prediction MAE minus the mean self-prediction MAE expressed as a percentage of the cross-prediction MAE, for each forecast decade from the 2030s to the 2090s. The self-prediction MAE compares each member's forecast to that member's true future climatology, and the cross-prediction MAE compares it to the true climatologies of all other members of the same ESM. The black line gives the relative MAE reduction computed from annually averaged errors, and the blue, green, red, and orange lines give the relative MAE reductions computed from December-February (DJF), March-May (MAM), June-August (JJA), and September-November (SON) seasonal-mean errors, respectively. Positive values indicate that forecasts are closer to the future of the member whose historical trajectory served as input than to the futures of the other members. The grey horizontal line marks zero reduction. ESMs used in this analysis are those with more than five members (Supplementary Table 1).

3.2.2 Spatial structure of skill from conditioning on recent trajectories

To attribute local forecast skill to the specific trajectory of internal variability, we next map the same relative MAE in reduction of each realisation within ESM ensembles at the grid-cell level for 11 of the 12 ESMs with more than five pure initial-condition members (Fig. 6). GISS-E2-1-G is excluded because its ensemble mixes initial-condition members with distinct physics and forcing variants (Miller et al. 2021), so cross-member differences are not attributable to internal variability alone (Section 3.2.1, Fig. 6).

While the U-Net produced significant but only modest skill improvements for global averages (Fig. 6), there are substantial improvements at the local level. In the 2030s the relative MAE reduction is significant over nearly the entire globe, strongest

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

over the subpolar North Atlantic, the mid-latitude Southern Ocean, and parts of the subtropical oceans (Fig. 7a). The relative MAE reduction then decreases gradually in most regions with lead time, and from the 2070s onward it becomes non-significant over much of the tropics, particularly the Indian Ocean and the tropical Pacific (Fig. 7b–g), where the signal-to-noise is low (King and Harrington 2018) and differ little across individual members (Fig. 1). The Southern Ocean retains coherent, significant MAE reductions at all lead times. The subpolar North Atlantic and adjacent Arctic and Nordic Seas stand out most clearly, with the strongest relative MAE reductions of any region at every forecast decade through the 2090s, maintaining a relative MAE reduction of 10% to 15% across future decades. Overall, the relative MAE reduction is remarkably one-sided: significantly positive over most of the globe at short range (2030s–2040s) and over large areas even at the century's end, and nowhere significantly negative at any lead time. The signal is therefore spatially-organised, varies regionally in magnitude, and is unambiguous in sign (Fig. 7).

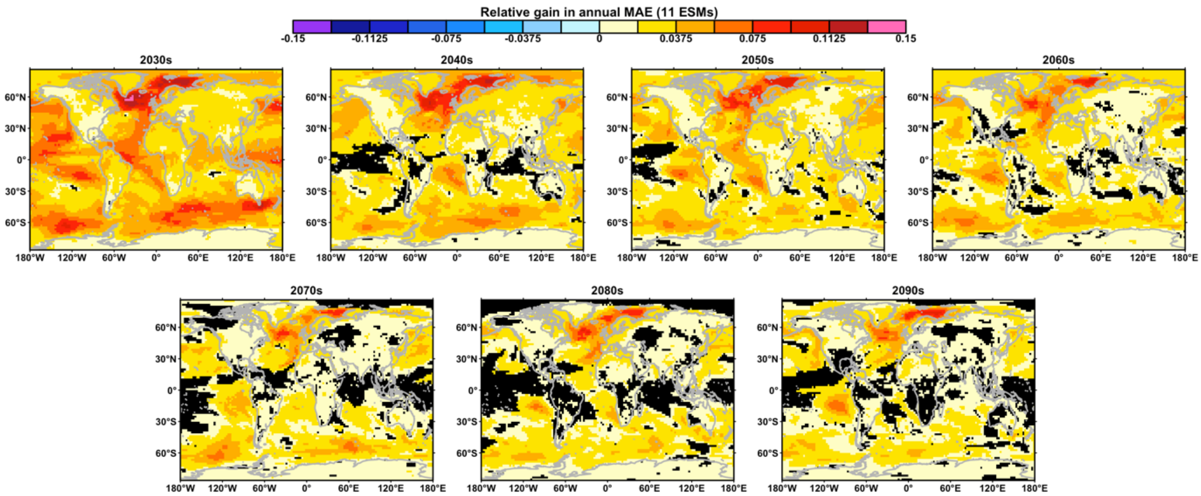


Figure 7: Spatial distribution of the relative MAE reduction from realisation-level information. Each panel maps, for one forecast decade from the 2030s to the 2090s, the relative MAE reductions in annual-mean MAE, defined as the cross-prediction MAE minus the self-prediction MAE expressed as a percentage of the cross-prediction MAE (Section 3.2.1), computed at each grid cell and pooled over the 11 ESMs with more than five ensemble members. Black shading masks grid cells where the relative MAE reduction is not significant at the 90% level. GISS-E2-1-G is excluded because its members span distinct physics and forcing variants rather than initial conditions alone (Miller et al. 2021).

The regions of highest relative MAE reductions are those where slow ocean dynamics carry multidecadal memory, the subpolar North Atlantic through AMOC and gyre heat-transport variability (Buckley and Marshall 2016, Zhang 2019) and the Southern Ocean through its long-timescale heat uptake and storage (Armour et al. 2016). This spatial pattern is consistent with initialised decadal prediction, whose added skill is largest and most robust in the subpolar North Atlantic and adjacent continents (Yeager and Robson 2017, Borchert et al. 2021). In addition, member-subselection approaches that extend initialisation information past the ten-year target of decadal prediction systems find their added value concentrated in the North Atlantic over roughly the first two decades (Befort et al. 2020, Mahmood et al. 2021, 2022). Our result shows an extension of the same signal to much longer range: trajectory-specific skill in the subpolar North Atlantic, and subsequently Europe and North America (Fig.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

7, Drews et al. 2024, Sutton and Dong 2012) remains detectable up to the 2090s. We assume this difference to be methodological capacity, since a deep network ingesting the full 46-year spatio-temporal trajectory can exploit the low-frequency ocean state nonlinearly, where subselection methods (e.g., Mahmood et al. 2021, 2022) reduce it to pattern agreement over a short matching window. The member-level design therefore reveals short- to long-range predictability that exists in the system.

3.3 Spatial structure of skill from conditioning on recent trajectories

To evaluate robustness beyond the CMIP6 training ensemble, we assess TC-UNet on simulations from the next-generation Met Office / Hadley Centre ESM HadGEM3-GC5 (hereafter GC5). GC5 succeeds the CMIP6 HadGEM3-GC3.1 family with substantial revisions to its dynamical core and physical parameterisations, informed by a decade of documented CMIP6 limitations in cloud and cloud-aerosol feedbacks (Zelinka et al. 2020, 2022, Meehl et al. 2020), ocean circulation (Jackson et al. 2022, Michel et al. 2025, Weijer et al. 2020), and regional pattern errors (Eyring et al. 2019, Sørland et al. 2018). It is being prepared for CMIP7 (Dunne et al. 2025) and therefore provides a test of generalisation under ESM changes (e.g. Brunner and Sippel 2023) that CMIP6-based LOMFO-CV alone cannot capture.
The GC5 suite used here contributes to the European Eddy-Rich ESM project, a grouped European contribution to HighResMIP2 for CMIP7 (Roberts et al. 2025), and additionally enables an assessment of cross-resolution generalisation. Single GC5 simulations under historical and SSP2-4.5 forcings have been carried out at three resolutions: a coarse configuration comparable to standard CMIP6 ESMs (GC5-LL; ~100 km ocean, ~250 km atmosphere), an eddy-permitting configuration similar to the previous HighResMIP generation (GC5-MM; ~25 km ocean, ~80 km atmosphere; Haarsma et al. 2016), and an eddy-rich configuration (GC5-HH; <10 km ocean and atmosphere; Hewitt et al. 2016, Roberts et al. 2025). GC5-HH represents a qualitative step change: mesoscale eddies on 10-100 km scales, central to ocean heat transport (Wang et al. 2023), mixing (Swingedouw et al. 2022), and air-sea coupling (Small et al. 2008), are explicitly resolved rather than parameterised. Because TC-UNet is trained exclusively on coarse-resolution ESMs, skill on GC5-HH would indicate that the learned relationships are robust to resolution rather than tied to coarse-grid parameterisation artefacts. This would build confidence in ultimately using TC-UNet to produce well-calibrated predictions for the real world.

3.3.1 Recovery of global mean trajectories across resolutions

We first evaluate how well TC-UNet recovers the GC5 decadal-mean GMST trajectory. For each target decade, we construct 5%–95% uncertainty ranges using the re-sampled empirical distribution (Section 2.2.3) of decadal GMST forecast errors obtained from the CMIP6 LOMFO-CV application (Supplementary Fig. 1). These uncertainty bounds quantify the expected forecast error when the network is applied to unseen ESMs, and hence includes ESM differences and realisation-specific errors (Supplementary Fig. 1). Fig. 8 shows the GMST values derived from full-field TC-UNet forecasts for GC5-LL, GC5-MM,

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

and GC5-HH, obtained from five-realisation ensemble-mean predictions (Section 2.2.2), with empirical uncertainty ranges subsequently computed via Monte-Carlo resampling (Section 2.2.3).

Across all resolutions and forecast horizons, the simulated GC5 GMST trajectories lie entirely within the TC-UNet uncertainty ranges (Fig. 8). The network not only reproduces the overall warming trajectory but also correctly captures the resolution dependence of GC5 warming: GC5-LL exhibits stronger warming than GC5-MM, which in turn warms more strongly than GC5-HH. This ordering is consistently reproduced from the 2030s through the 2090s. The distinct warming rates recovered using only the preceding 46 years of temperature anomalies (relative to 1995–2014) indicates that TC-UNet extracts information from recent variability and trends that is predictive of resolution-dependent forced responses (through differences in parametrised and explicitly resolved physics), rather than relying on features specific to the coarse-resolution training ensemble.

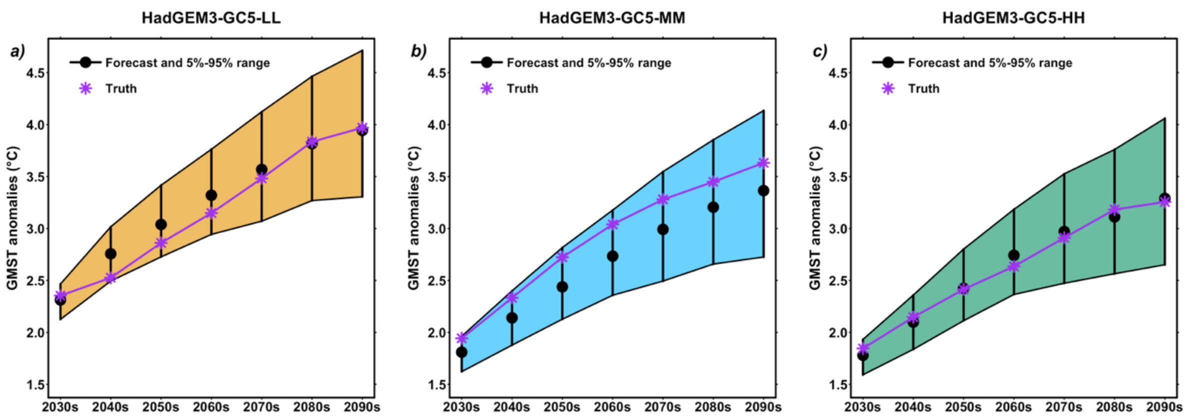


Figure 8: True and predicted global mean surface temperature (GMST) anomalies for HadGEM3-GC5 simulations. Results are represented for HadGEM3-GC5-LL (a), HadGEM3-GC5-MM (b), and HadGEM3-GC5-HH (c). The forecasts are represent by black dots while true GMST values are represented by purple asterisks. The forecast uncertainty range is determined from empirical errors (Supplementary Fig. 1) and indicated by horizontal black lines and yellow (a, GC5-LL), blue (b, GC5-MM), and green (c, GC5-HH) areas.

For GC5-LL, the simulated warming in the 2040s reaches the lower bound of the 90 % uncertainty interval prediction from TC-UNet (Fig. 8a). This behaviour arises because the simulated warming between the 2030s and 2040s in GC5-LL is anomalously weak relative to both GC5-MM and GC5-HH, despite GC5-LL exhibiting higher short- and long-term mean warming. This pattern is consistent with a specific realisation of decadal-scale internal variability that temporarily attenuates the forced warming signal (e.g., Cassou et al. 2018, Kosaka and Xie 2013). Importantly, this outcome remains within the empirically derived uncertainty range which accounts for such shifts by construction (i.e., −0.23 °C within [−0.26 °C, 0.26 °C], Supplementary Fig. 1, Fig. 8a, Section 2.2.2). From the 2050s onward, the mean TC-UNet prediction and the simulated GC5-LL trajectories gradually converge, supporting the interpretation that the early-decade discrepancies (2040s–2060s) primarily reflect a warming slowdown induced by internal variability rather than a systematic ESM mismatch.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

GC5-MM illustrates a complementary behaviour, in which we observe a systematic ESM-specific offset that is also encompassed by the uncertainty framework (Fig. 8b). Across decades, the true GMST in GC5-MM is consistently warmer than the TC-UNet ensemble-mean prediction by approximately 0.13 °C in the 2030s, increasing to about 0.31 °C by the 2060s. All such offsets fall comfortably within the 90% uncertainty range derived from resampled LOMFO-CV errors (Fig. 8), demonstrating that the empirical uncertainty estimates successfully account for both internal variability shifts (GC5-LL, Fig. 8a) and specific ESM shifts GC5-MM, Fig. 8b).

Finally, TC-UNet yields the most accurate GMST forecasts for GC5-HH in terms of mean predictions across the GC5 hierarchy (Fig. 8c). Errors in decadal-mean GMST range from −0.11 °C to 0.07 °C across all lead times, representing remarkably high accuracy from the near-term to the end-of-century horizon. This result provides strong evidence that TC-UNet does not rely on artefacts of CMIP6 parameterisations but instead captures robust relationships between recent large-scale spatiotemporal temperature evolution and future climatological states. The observation that training exclusively on low-resolution CMIP6 ESMs suffices to reproduce the behaviour (at this level, for the extracted GMST) of an eddy-rich ESM configuration substantially strengthens confidence in the applicability of the TC-UNet framework to more realistic ESMs in terms of explicitly resolved physics, and, by extension, to real-world observations.

3.3.2 Spatial structure of forecast errors in GC5 configurations

We next assess the generalisability of TC-UNet to unseen and higher-resolution physics at the grid-cell level, extending the global-mean analysis of Section 3.2.1 to spatially resolved fields. The following spatial analyses are restricted to annually averaged TC-UNet forecasts and GC5 projections. For each grid cell, empirical 5%–95% uncertainty ranges are constructed from out-of-sample LOMFO-CV forecast errors, analogously to the GMST uncertainty estimation (e.g., Supplementary Fig. 1, Section 2.2.3) but applied locally rather than globally.

Fig. 9a–g shows all decadal-mean surface air temperature forecast-error maps for the GC5-HH configuration only, with shading highlighting grid cells where obtained errors fall outside the grid-cell-specific uncertainty ranges. This analysis demonstrates that the strong skill identified for GC5-HH at the global-mean temperature level is mirrored spatially. Across decades, between 93.3% (2030s, Fig. 9d) and 99% (2050s, Fig. 9c) of the global surface area lies within the local 90% empirical uncertainty intervals.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

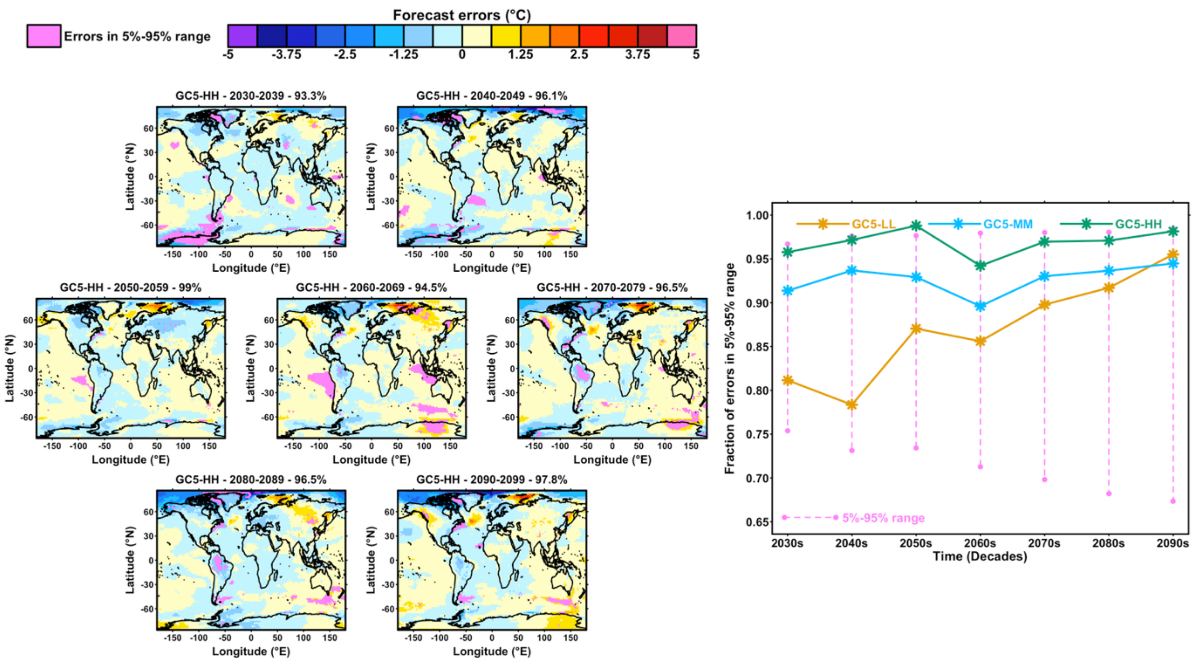


Figure 9: True and predicted local temperature anomalies for HadGEM3-GC5-HH. Results are represented for HadGEM3-GC5-HH spatially, for predicted and true value averaged across calendar months. On the left hand side (maps), each forecast decade from 2030s to 2090s is represented by the different panels, from left to right and top to bottom. For each map, the colours represented in the colour bar at the top indicate the forecast errors in °C. The orchid-coloured areas indicate where the forecast errors are outside the empirically determined error bars from the leave-one-model-family-out cross-validation (LOMFO-CV). On the right-hand side the panel indicate the fraction of the total area where errors are outside of the empirically determined error bars from the LOMFO-CV. The green, blue, and yellow lines indicate coverage values for HadGEM3-GC5-HH, HadGEM3-GC5-MM, and HadGEM3-GC5-LL, respectively. The orchid dots connected by vertical dashed lines represent the coverage values as obtained through the LOMFO-CV (where each model family's coverage is determined by the LOMFO-CV errors from all other model families).

The spatial structure of out-of-range grid cells reveals physically interpretable patterns rather than widespread forecast degradation. These patterns can be broadly classified into three categories:

(i) Statistically expected occurrences, consisting of small, spatially scattered exceedances consistent with noise and residual internal variability (e.g. Fig. 9a–g).

(ii) Physically coherent deviations associated with large scale internal variability modes, such as an Indian Ocean Dipole-like (e.g., Saji et al. 1999) pattern emerging in the 2060s, and a bidecadal eastern tropical Pacific error structure in the 2050s–2060s, the latter being consistent with known interdecadal to multidecadal Pacific variability linked to variations in surface upwelling strength in this particular covered region (e.g., Power et al. 2021).

(iii) Regionally persistent deviations associated with out-of-sample ESM behaviour, most notably along the Gulf Stream path (2040s–2090s; Fig. 9b–g) and in the Labrador Sea (2030s–2040s and 2080s–2090s; Fig. 9a,b,g–h) for GC5-HH, both regions characterised by intense mesoscale eddy activity (Roberts et al. 2016, 2020, Swingedouw et al. 2022).

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

None of these error structures indicate a globally degraded forecast skill for GC5-HH. Instead, they reflect localised, physically interpretable, and somewhat expected shifts within an otherwise well-calibrated spatial prediction framework (Fig. 9).

Because forecast errors across grid cells are spatially correlated, exceedances of local uncertainty ranges are not expected to occur independently. Consequently, the nominal 5%–95% uncertainty range applies strictly at the grid-cell level by construction but does not imply that exactly 10% of the global area for a given predicted climatology should lie outside these bounds. To account for this spatial dependence, for each decade we derive from the LOMFO-CV experiments the empirical 5%–95% range of the area-weighted fraction of grid cells whose forecast errors fall outside their local uncertainty intervals. This defines an expected envelope for the fractional coverage of local forecasts falling in the 5%–95% confidence interval (Fig. 9h).

Across all three GC5 configurations, the observed fractional coverage of grid cells lying within their local 5%–95% uncertainty ranges falls within this empirically derived envelope for 20 out of 21 forecasts (i.e., seven decades across three configurations, Fig. 9h). The only exception occurs for GC5-HH in the 2050s, where the fraction of in-range grid cells exceeds the 95th percentile of the LOMFO-CV distribution, indicating high spatial skill relative to what is inferred from CMIP6-only validation (Section 2.2.3). Consistent with the global-mean analysis, the comparatively lower short-range skill of GC5-LL is also evident at the spatial level (Fig. 9h).

At first glance, the superior spatial skill of GC5-HH relative to GC5-LL may appear counterintuitive, given that GC5-LL shares a resolution regime closer to that of the training ensemble. This difference may partly reflect specific realisations of internal variability as hypothesised from GMST forecasts (Fig. 8), though this cannot be robustly assessed given the availability of only a single member per GC5 configuration. An additional explanation is that forecast skill is mainly influenced by the magnitude of ESM warming rather than resolution alone. GC5-LL exhibits particularly strong warming (e.g. +3.90 °C for 2080–2099 relative to 1850–1900, Fig. 8a), placing it near the upper tail of the CMIP6 warming distribution from our multi-model and multi-member ensemble (+4.03 °C at the 95th percentile, Table 1), and therefore in a regime that is sparsely represented in the TC-UNet training space. By contrast, GC5-HH exhibits substantially lower late-century warming (+3.22 °C for 2080–2099), placing it closer to the centre of the training distribution, with a median warming of +2.89 °C for 2080–2099. Overall, this analysis indicates that the TC-UNet performance when applied to a more realistic physical system is only weakly affected by the highly parameterised nature of the training ensemble.

In line with the need for out-of-sample validation in such uncertainty reduction exercise (Nowack and Watson-Parris 2025), we argue that such cross-resolution validation is a strong applicability check when transferring learned relationships from coarse-resolution ESMs to observations. This is particularly valid as high-resolution climate simulations become increasingly available (Haarsma et al. 2016, Roberts et al. 2025) but remain too limited in number for their own projections to be independently constrained.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

3.4 Real-world future forecasts

Following the validations of Sections 3.1 and 3.2, we apply the framework to ERA5 and BEST. Notably, the network is trained on surface air temperature, which ERA5 provides consistently over land and ocean, whereas BEST blends air temperature over land with sea surface temperature over ocean, which warms more slowly, so the network receives abnormally weak oceanic trends (Section 2.1.2). Conversely, IPCC warming levels are defined for blended observational data, so ERA5-based forecast anomalies are expected to be slightly warmer on the global average because they use air temperature rather than sea surface temperature over the ocean (Section 2.1.2). BEST serves as a robustness check against other GMST constraints (Lee et al. 2021, Ribes et al. 2021).

3.4.1 Global mean surface temperature and uncertainty

We first evaluate the framework on global mean temperature, the high signal-to-noise quantity for which observational constraints are best established and against which our forecast can be compared directly. For each target decade we apply the network to the ERA5 and BEST inputs and derive 5%–95% uncertainty ranges from the CMIP6 LOMFO-CV error distribution (Section 2.2.3), so that the reported spread reflects performance on structurally unseen ESMs and retains both inter-ESM differences and internal variability.

As developed in Section 2.1.2., we recall that the two real-world data products probe different aspects of the framework because they differ in physical content. Forced with BEST, the TC-UNet forecast for 2080–2099 is 2.64°C (5%–95% interval: 2.06 to 3.35°C; Fig. 10a), in close agreement with the central estimates of the observationally constrained range of Ribes et al. (2021) (2.68°C) and the IPCC AR6 assessed range (2.65°C, Lee et al. 2021) for SSP2-4.5 (Fig. 10a). This agreement rests on a consistent observational basis: BEST blends land air temperature with sea surface temperature, matching the construction of the GMST records on which these constraints are built. Forced with ERA5 2-metre temperature, the forecast is warmer, 2.97°C (5%–95% interval: 2.39 to 3.68°C; Fig. 10b), as expected for a GSAT-type quantity that samples air temperature rather than sea surface temperature over the ocean.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

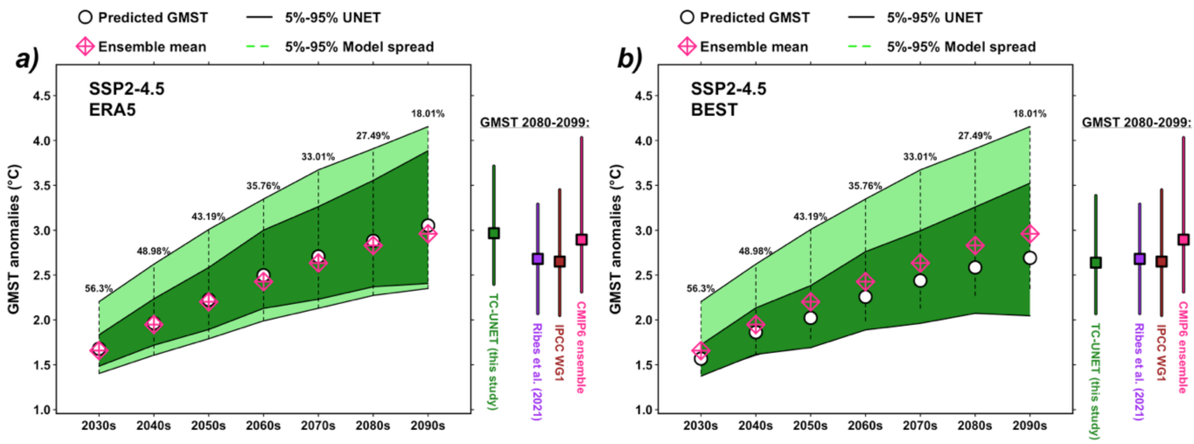


Figure 10: Projected global mean surface temperature (GMST) under SSP2-4.5 from ERA5 (a) and BEST (b). White markers show the TC-UNet decadal-mean GMST forecast, with the dark-shaded band the 5%–95% uncertainty range from the LOMFO-CV error distribution and the light-shaded band the 5%–95% CMIP6 Earth System Model (ESM) spread. Crossed pink diamonds show the CMIP6 ensemble mean. Percentages above each decade give the reduction in the TC-UNet range relative to the ESM spread. Anomalies are relative to the 1995–2014 climatology. The bars at the right of the panels compare the 2080–2099 the GMST 5%–95% uncertainty from TC-UNet (this study), Ribes et al. (2021, purple), IPCC AR6 WG1 (Lee et al. 2021, brown), and the CMIP6 ensemble (pink). The associated squares indicate central estimates.

The forecast uncertainty range of 1.29°C (Fig. 10) is comparable to that of Ribes et al. (2021) (1.22°C) and the AR6 assessed range (1.4°C, Lee et al. 2021), and this equivalence is reached under stricter conditions. The Bayesian method underlying Ribes et al. (2021) targets the forced response: observed internal variability is modelled explicitly and filtered from the constraint, so the resulting interval excludes the variability of the single observed realisation. Our interval is instead constructed from empirical out-of-sample forecast errors at the level of individual realisations, and therefore retains the residual internal variability that no single realisation allows one to remove, a contribution that Li et al. (2025) show propagates measurably into constrained predictions of GMST. It is also derived from a larger and more structurally diverse ensemble (40 ESMs against 27), which alters the error distribution the forecast is evaluated against. Matching the constrained global ranges while carrying additional sources of uncertainty indicates that TC-UNet extracts meaningful information from the recent observed trajectory, in line with the prediction framings of Li et al. (2025) and Li et al. (2026). The comparison also understates the difference in scope: the constraints we compare to are scalar projections of global mean warming. TC-UNet predicts maps of decadally-averaged monthly mean temperature fields, and the global result discussed here is a single aggregated diagnostic of that field. Reaching parity on the quantity these methods are optimised for, while resolving the local and seasonal structure they do not predict, is a central added value of the approach.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

3.4.2 Spatial distribution of uncertainty reductions

We next extend the analysis to the grid-cell level, computing local forecasts and 5%–95% ranges as for GMST but from the resampled local LOMFO-CV error distribution. To quantify the gain relative to the raw ensemble we define the spread-reduction ratio

$$R = 1 - \frac{CI_{(5-95\%,LOMFO-CV)}}{CI_{(5-95\%,ENSEMBLE)}} \qquad (8)$$

so that positive R denotes a narrower forecast interval than the unconstrained ensemble; R is computed from the resampled range (Section 2.2.3).

Fig. 11 shows the spatial distribution of R for short- (2030s–2040s), medium- (2050s–2070s), and long-range (2080s–2090s) forecasts. The network narrows the uncertainty range across almost the entire globe at every horizon, reaching up to 75% for short- and medium-range local forecasts, and 60% for long-term ones. The only exceptions are a few small long-range areas over parts of the North Pacific, Texas, and the Iberian Peninsula, where the interval is left essentially unconstrained. Median reductions are 39%, 31%, and 22% for the short, medium, and long-range.

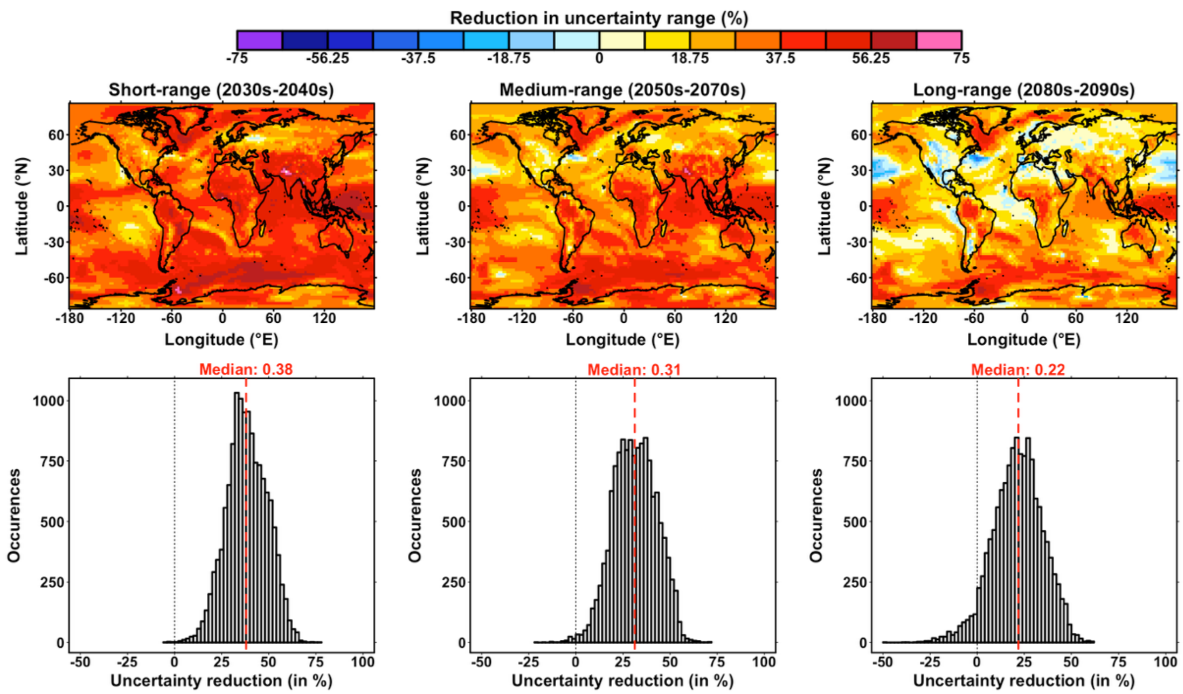


Figure 11: Local spread reduction. Local reduction in forecast uncertainty under SSP2-4.5. a–c): maps of the grid-cell reduction in the 5%–95% uncertainty range relative to the unconstrained ensemble, $R = 1 - \frac{CI^{LOMFO-CV}_{5\%-95\%}}{CI^{ENSEMBLE}_{5\%-95\%}}$, as a percentage, for short- (2030s–2040s), medium- (2050s–2070s), and long-range (2080s–2090s) forecasts, respectively. Positive values (warm colours) denote a narrower TC-UNet interval than the raw ensemble. d–f) distributions of the grid-cell reductions for each range, for short- (2030s–2040s), medium- (2050s–2070s), and long-range (2080s–2090s) forecasts, respectively. Medians of distributions are indicated by the vertical dashed red lines.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

These reductions act on a different quantity than the local constraints of Qasmi and Ribes (2022), who report 30–70% for SSP5-8.5. Their prior is built from 27 ESMs with internal variability filtered out, so the constraint narrows forced-response ESM uncertainty alone. Ours is built from 40 ESMs and retains internal-variability realisations, which at the local scale and over multi-decadal horizons constitute a large, irreducible fraction of total uncertainty (Hawkins and Sutton 2009, Lehner et al. 2020, Deser 2020, Deser et al. 2020). A given percentage therefore corresponds to a larger absolute narrowing in our framework, and the two sets of values track progress against different baselines. The approaches also differ structurally: local Bayesian constraint is applied independently at each grid cell, whereas TC-UNet reconstructs the full spatial map through its U-Net decoder (Fig. 2), and the smooth, physically consistent structure of R in Fig. 11 is the direct expression of this design.

3.4.3 Regional-scale projections and uncertainties

Because the network forecasts the full spatial field, regional means can be aggregated directly from the gridded output. We compute decadal means over eight sub-continental regions from the ERA5-based forecasts (Fig. 12), with 5%–95% ranges propagated from the local LOMFO-CV errors as in Section 3.4.2. The regional forecasts depart from the raw ensemble mean in a spatially structured way that does not reduce to a single sign. Over Europe and North Asia the forecast tracks at or below the ensemble mean across all decades, indicating constrained warming weaker than the unconstrained CMIP6 average. For South America, forecasts generally match the ensemble mean or slightly above. In the remaining regions, forecasts' central estimates remain broadly consistent with the ensemble mean.

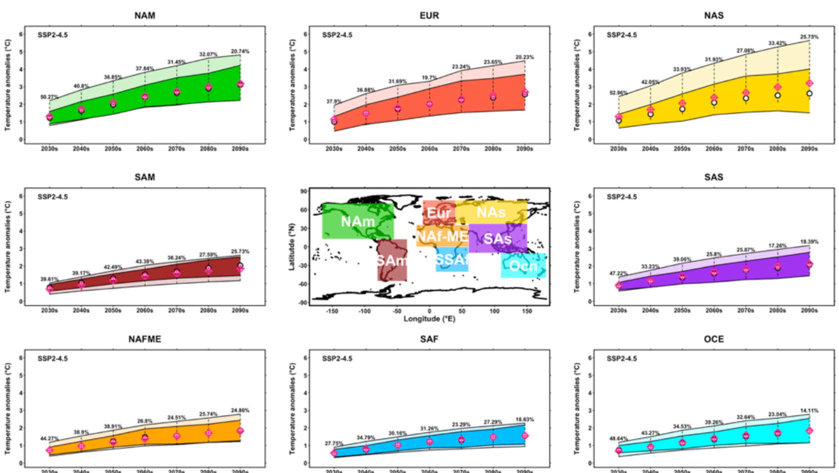


Figure 12: Regional to continental scale forecasts from ERA5. Each panel shows the TC-UNet forecast of decadal-mean surface air temperature anomaly for one of eight sub-continental regions, with the shaded band giving the 5%–95% uncertainty range from the LOMFO-CV error distribution, for each target decade from the 2030s to the 2090s. White circles show the predicted ERA5 trajectory and pink crossed diamonds the CMIP6 ensemble mean. Percentages above each envelope give the reduction in spread relative to the unconstrained ensemble. Regions are defined in the central map: North America (NAM, green), Europe (Eur, red), North Asia (NAs, yellow), South America (Sam, brown); North Africa and Middle East (NAf-ME, orange); South Asia (Sas, purple); sub-Saharan Africa (SSAF, blue), Oceania (Oce, cyan). Anomalies are relative to the 1995–2014 climatology.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

The spatial heterogeneity of regional deviations with the ensemble mean is informative because it coexists with near-exact agreement at the global scale (Fig. 10): a globally near-neutral adjustment results in compensating regional departures from the ensemble mean, which only a spatially resolved forecast can reveal. In summary, by forecasting the spatial field directly, TC-UNet resolves continental-scale differences in constrained warming while preserving global consistency, the regime most relevant for adaptation planning.

3.4.3 Discussion

This work shows that deep learning can contribute meaningfully to uncertainty reduction in climate projections, alongside a recent move toward constraining future responses with learned rather than prescribed relationships (Ricard et al. 2024, Immorlano et al. 2025, Li et al. 2026, Portmann et al. 2026). We treat the diversity of ESMs, and of per-ESM realisations where available, as high-dimensional information linking the recent climate trajectory to the future decadal state. Unlike observational constraints, the method does not depend on a present-to-future relationship between compressed scalar metrics such as equilibrium climate sensitivity (e.g., Cox et al. 2018, Ricard et al. 2024), which can hold by chance (Caldwell et al. 2014, Nowack and Watson-Parris 2025) or through shared ESM biases (Sanderson et al. 2021). It also carries internal-variability realisations explicitly, in both the prior and the posterior, and this property shapes the results that follow.
Retaining internal variability sets the scale dependence of the uncertainty reduction. For end-of-century GMST, where internal variability is small relative to the forced signal (Hawkins and Sutton 2009), the method reaches a reduction comparable to Bayesian constraints (Ribes et al. 2021, 2025), even though global mean temperature is not its target, but a quantity derived from a much richer spatial forecast. At that scale, both approaches primarily draw their reduction from ESM diversity. At the local scale the reduction is smaller than for Bayesian constraints (Fig. 11, Qasmi and Ribes 2022), because in many locations internal variability is a large fraction of total uncertainty (Fig. 1) and our intervals keep it rather than filtering it out. The two cannot be compared directly: our ranges describe the realised future state, including the variability that a single observed trajectory cannot remove, whereas constraint methods describe the de-noised forced response alone.
We also find that TC-UNet forecasts are systematically closer to the future of the member whose trajectory they received than to the futures of other members of the same ESM (Figs. 6–7). The network therefore extracts predictive information from the specific realisation of low-frequency variability carried in the input, not only from ESM identity. This trajectory-specific skill is spatially organised, strongest and most persistent over the subpolar North Atlantic, where slow ocean dynamics provide multidecadal memory (Buckley and Marshall 2016, Zhang and Sun 2019), and it remains detectable to the 2090s (Figs. 6–7). That is well beyond the ten-year horizon of initialised prediction systems (Boer et al. 2016, Yeager and Robson 2017, Borchert et al. 2021) and the roughly two decades over which member-subselection methods retain added value (Befort et al. 2020, Mahmood et al. 2021, 2022). Trajectory-specific predictability therefore persists at multidecadal range precisely where projection uncertainty from internal variability is largest (Gu et al. 2024). This suggests that the observed trajectory carries information capable of narrowing the wide range of possible North Atlantic mid- to long-term futures beyond model

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

uncertainty (Bonan et al. 2025, Portmann et al. 2026), and potentially of the AMOC evolution underlying it (Weijer et al. 2020, Jackson et al. 2022, Michel et al. 2025).

The credibility of these results rests on the validation framework, itself a central contribution of our work. Uncertainty-reduction methods are hard to evaluate because the future they target is unobserved, and in-sample skill is a weak guide to real-world performance (Nowack and Watson-Parris 2025). We propose LOMFO-CV, which tests every forecast against an ESM family withheld in full, so skill is measured on unseen physics rather than on the ensemble used for training. Grouping by family rather than by individual ESM is essential here, because ESMs from the same centre often share dynamical cores, coupling strategies, and parameterisations (Brunner et al. 2020, Brunner and Sippel 2023, Knutti et al. 2017, Michel et al. in rev.). The resulting error distributions reflect generalisation across ESM structure, and the uncertainty bounds are calibrated against that generalisation rather than fitted in-sample.

The framework extends naturally to validation against entirely new ESM generations. Applying the network to the HadGEM3-GC5 hierarchy, including its eddy-rich configuration, tests whether relationships learned on coarse CMIP6 physics survive a step change in resolved processes outside the training distribution. This test matters increasingly: high-resolution ESMs are emerging (Haarsma et al. 2016, Roberts et al. 2025) but will long remain too few and too expensive to form ensembles large enough to constrain their own projection uncertainty. A method that learns from the large pool of coarse-resolution simulations, and is explicitly validated via transfer to higher-resolution models with richer explicit physics, offers a practical route to validating constrained projections and climate prediction.

The uncertainty estimates carry caveats that must be explicitly identified. First, the empirical ranges are limited by the uneven composition of the CMIP6 ensemble: member counts range from one to fifty across the 40 ESMs (Supplementary Table 1), so internal variability is richly sampled for a few ESMs and barely for many, and the Monte Carlo resampling that balances ESM weights (Section 2.2.3) is a pragmatic correction rather than a learned quantity. A natural improvement is to learn the predictive distribution during training through heteroscedastic likelihoods, for example a Gaussian negative-log-likelihood loss (Lakshminarayanan et al. 2017) for local temperature. Such probabilistic losses would remove the resampling step and let the network attribute uncertainty itself, in particular widening intervals where the input trajectory departs from anything seen in training. The data side should also improve: the CMIP7 ScenarioMIP protocol encourages ensembles of ideally five or more members per scenario (van Vuuren et al. 2026), and applying the framework to such more balanced ensembles would strengthen both the training signal and the empirical error estimation. Second, the cross-resolution evaluation rests on a single member per GC5 configuration, so the agreement we report reflects one realisation of internal variability; larger high-resolution ensembles, or simulations from other high-resolution ESMs, will be needed to confirm it. None of these caveats undermines the central results of this work.

The same retention of internal variability that limits the local reduction is what makes the framework most promising where constraints are hardest. Because it forecasts future states directly rather than de-noised forced responses, it is naturally suited to low signal-to-noise variables (Fig. 1) and could be tested on quantities that resist conventional constraint. Whether the

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

framework is able to reduce uncertainty in quantities where internal variability contributes more strongly, such as regional precipitation (O'Reilly et al. 2024), or change in compound extremes (Bevacqua et al. 2022, 2023), has yet to be tested. Observationally constrained projections have been argued to be predictions rather than projections, but only at the global scale, where the signal rises clearly above internal variability (Li et al. 2025). At the regional and local scales that matter for adaptation, the single realisation we observe carries internal variability that constraint methods discard by construction. By learning directly from the recent observed trajectory and keeping that variability in calibrated out-of-sample uncertainty ranges, the framework presented here extends the prediction framing from the global mean to the spatial field. The result is a forecast of future decadal climate that is spatially resolved and interpretable in the source of its uncertainty reduction (Figs. 6–7). As the ESM landscape grows richer (Dunne et al. 2025) and higher in resolution (Roberts et al. 2025), the value of recent observations for conditioning the decades ahead will only increase. Methods that read that information directly from the trajectory, rather than from a handful of summary metrics, are then central to turning climate projections into actionable climate predictions, even more so because the past and future climate experienced on Earth is a single realisation.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Appendix A

In this appendix, we formally describe the neural network (NN) architecture, identify its trainable parameters, and outline the optimisation strategy used during training. The description is general and applies to any combination of SSP scenario and target decade. Conceptually, the NN is composed of two main components: (i) a grid-cell–wise temporal encoder, which extracts multi-scale temporal features from the recent climate evolution, and (ii) a U-Net spatial decoder, which operates on the resulting latent representation to reconstruct the future monthly climatology. These two modules together form a temporal-to-spatial prediction framework tailored to the problem of future climate-field projection from observed variations.

A.1 Temporal feature extraction

The input to the network is a sequence of monthly surface air temperature anomalies with dimensions (N, T, H, W), where N is the number of training samples, T the number of time steps, and $H \times W$ the latitude-longitude grid (here with $H = 180$ and $W = 360$, see section 2.1). This input is converted to a 5-D tensor $(N, T, H, W, 1)$ by adding a singleton channel dimension, necessary for the NN training.

To reduce computational cost and focus on relatively low frequencies, we apply two consecutive 3-D average pooling layers along the time axis only, which is achieved by a pool size and stride of $(4,1,1)$. This compresses the temporal dimension by a factor of approximately 16 (from 372 to 24) while leaving the spatial resolution unchanged.

The pooled sequence is passed through a series of 3-D temporal convolutional blocks that operate only along the time dimensions which also only operate along the time direction. Each block consists of 3-D convolution with kernel size $(5,1,1)$ and a 3-D convolution with kernel size $(3,1,1)$, both with a GELU activation, followd by batch normalisation. We use three such blocks with increasing temporal dilation rates $d = 1, 2, 4$. Time-dilated convolutions extend the temporal range of each kernel by skipping over intermediate time steps, so the kernel samples points that are d steps apart rather than adjacent ones. The dilation factor d therefore controls how stretched the kernel is along the compressed time axis, enlarging the receptive field without increasing the number of parameters. Hence with the prior time compression and $d = 1, 2, 4$, the encoder simultaneously captures pluriannual variability ($d = 1$), decadal variability ($d = 2$), and longer term forced trends ($d = 4$). Because the kernels are of size $(k, 1,1)$, these convolutions act as temporal filters applied identically at each grid point, with shared trainable parameters across all lat-lon locations. This makes each latent channel represent a learned temporal pattern that is consistant in space.

Finally, the temporally filtered tensor is projected to a latent feature space via a $1 \times 1 \times 1$ 3-D convolution with D_{latent} output channels, here with $D_{latent} = 24$, and then averaged over the compressed time dimension. This yields a latent representation with shape (N, H, W, D_{latent}), *i.e.* a set of D_{latent} temporal-encoder feature maps defined on the full spatial grid (H, W). All the parameters in the temporal encoder (3-D convolution kernels and biases, batch-normalisation scale and shift) are trained

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

and shared uniformly across all spatial locations, ensuring that the resulting feature maps represent a spatially consistent encoding of the temporal evolution.

A.2 U-Net

The latent tensor (N, H, W, D_{latent}) obtained from the temporal encoder serves as input to a 2-D U-Net based spatial decoder. Because the latitude dimension H is not an explicit multiple of 8 we first apply a zero-padding along longitude and latitude such that H an W are divisible by 8, ensuring that three successive 2×2 max-pooling operations can be performed without shape mismatches

The spatial decoder follows a standard encoder-decoder U-Net structure with three resolution layers. Each level uses a whole convolutional block consisting of two 3×3 2-D convolutional layers with GELU activation, followed by batch normalisation and a 10% dropout. The number of filters at each level is given by $\{24,48,96\}$. In the encoder path, each convolutional block is followed by a 2×2 max-pooling operation, progressively coarsening the spatial resolution and increasing the channel depth. The final convolutional block forms the U-Net bottleneck: a highly compressed spatial representation that also carries the temporal information encoded upstream (A.1).

In the decoder path, spatial resolution is restored using 2-D transposed convolutional layers with 2×2 kernels and stride, which perform the upsampling toward higher (initial) spatial resolution. Each of these are followed by standard 3×3 convolutional blocks (as in the U-Net encoder path) that refine the upsampled features. At each spatial scale of the deconvolution path, the upsampled feature maps are concatenated with the corresponding encoder features via skip connections. These skip connections enable the NN to combine coarse-scale information about large-scale patterns from the upsampled maps with fine-scale regional structure preserved in the downsampled U-Net's encoder features.

A final 1×1 2-D convolutional layer maps the output of the last decoder block to 12 channels corresponding to the predicted monthly climatology of the target decada (one channel per calendar month). All kernels and biases from 2-D convolutional and transposed convolutional layers, along with the bath normalisation parameters from the downsampling U-Net's branch, are trainable.

In summary, the architecture can be described as a pluriannual to multi-decadal temporal encoder - spatial U-Net decoder: the first stage compresses each grid cell's recent temporal evolution into a shared set of latent channels, and the second stage uses those latent spatial maps to reconstruct the decadal-mean seasonal cycle of future temperature at each spatial location.

A.3 Training procedure and optimisation

For each SSP–decade combination, a separate instance of the network is trained. Inputs are 31-year sequences of monthly temperature anomalies, and targets are 10-year monthly climatologies for the corresponding future decade, both defined on

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

the common $1° \times 1°$ grid. The data are split into training and validation sets by model/member, and batches are drown along the sample dimension N.

Including both the temporal encoder (A.1), and the U-Net (A.2), the NN contains roughly 3.762 million trainable parameters

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Code availability

The codes developed for this study are publicly available on Zenodo: https://zenodo.org/records/22392245

Data availability

The CMIP6 data used in this study are publicly available through the Earth System Grid Federation (ESGF). The data analysed here were downloaded from the ESGF-West node: https://metagrid.esgf-west.org.
The Berkeley Earth Surface Temperature (BEST) dataset is publicly available at https://berkeleyearth.org/data/.
The ERA5 monthly averaged data on single levels are publicly available through the Copernicus Climate Data Store at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means.
The output data analysed in this study are publicly available on Zenodo: https://zenodo.org/records/22392245

Author contributions

SLLM led and designed the study, pre-processed the data, carried out the analysis, and was the main writer of the manuscript. HMC contributed to the design og the study and the analysis. KS and HMC contributed to the manuscript writing and results assessment.

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Acknowledgements

The authors acknowledge the use of computing capacity from the DKRZ Levante supercomputer. The authors used AI tools for coding assistance and polish the manuscript.

Financial support

SLLM, KS and HMC acknowledge funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10049639] for the University of Oxford's contribution to the EERIE project. This work was part of the EERIE project which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101081383.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

References

Armour, K. C., Marshall, J., Scott, J. R., Donohoe, A., and Newsom, E. R.: Southern Ocean warming delayed by circumpolar upwelling and equatorward transport, Nat. Geosci., 9, 549-554, https://doi.org/10.1038/ngeo2731, 2016.

Befort, D. J., O'Reilly, C. H., and Weisheimer, A.: Constraining projections using decadal predictions, Geophys. Res. Lett., 47, e2020GL087900, https://doi.org/10.1029/2020GL087900, 2020.

Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., et al.: Advancing research on compound weather and climate events via large ensemble model simulations, Nat. Commun., 14, 2145, https://doi.org/10.1038/s41467-023-37847-5, 2023.

Bevacqua, E., Zappa, G., Lehner, F., and Zscheischler, J.: Precipitation trends determine future occurrences of compound hot-dry events, Nat. Clim. Change, 12, 350-355, https://doi.org/10.1038/s41558-022-01309-5, 2022.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., et al.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751-3777, https://doi.org/10.5194/gmd-9-3751-2016, 2016.

Bonan, D. B., Thompson, A. F., Schneider, T., Zanna, L., Armour, K. C., et al.: Observational constraints imply limited future Atlantic meridional overturning circulation weakening, Nat. Geosci., 18, 479-487, https://doi.org/10.1038/s41561-025-01709-0, 2025.

Borchert, L. F., Koul, V., Menary, M. B., Befort, D. J., Swingedouw, D., Sgubin, G., and Mignot, J.: Skillful decadal prediction of unforced southern European summer temperature variations, Environ. Res. Lett., 16, 104017, https://doi.org/10.1088/1748-9326/ac20f5, 2021.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., et al.: Comparing methods to constrain future European climate projections using a consistent framework, J. Climate, 33, 8671-8692, https://doi.org/10.1175/JCLI-D-19-0953.1, 2020.

Brunner, L. and Sippel, S.: Identifying climate models based on their daily output using machine learning, Environmental Data Science, 2, e22, https://doi.org/10.1017/eds.2023.23, 2023.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Buckley, M. W. and Marshall, J.: Observations, inferences, and mechanisms of the Atlantic Meridional Overturning Circulation: A review, Rev. Geophys., 54, 5-63, https://doi.org/10.1002/2015RG000493, 2016.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., et al.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophys. Res. Lett., 41, 1803-1808, https://doi.org/10.1002/2014GL059205, 2014.

Cassou, C. and Cattiaux, J.: Disruption of the European climate seasonal clock in a warming world, Nat. Clim. Change, 6, 589-594, https://doi.org/10.1038/nclimate2969, 2016.

Cassou, C., Kushnir, Y., Hawkins, E., Pirani, A., Kucharski, F., Kang, I.-S., and Caltabiano, N.: Decadal Climate Variability and Predictability: Challenges and Opportunities, B. Am. Meteorol. Soc., 99, 479-490, https://doi.org/10.1175/BAMS-D-16-0286.1, 2018.

Cox, P. M., Huntingford, C., and Williamson, M. S.: Emergent constraint on equilibrium climate sensitivity from global temperature variability, Nature, 553, 319-322, https://doi.org/10.1038/nature25450, 2018.

Deser, C.: Certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management, Earth's Future, 8, e2020EF001854, https://doi.org/10.1029/2020EF001854, 2020.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., et al.: Insights from Earth system model initial-condition large ensembles and future prospects, Nat. Clim. Change, 10, 277-286, https://doi.org/10.1038/s41558-020-0731-2, 2020.

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., et al.: Intermodel spread in the pattern effect and its contribution to climate sensitivity in CMIP5 and CMIP6 models, J. Climate, 33, 7755-7775, https://doi.org/10.1175/JCLI-D-19-1011.1, 2020.

Drews, A., Schmith, T., Tian, T., Wang, Y., Devilliers, M., et al.: The Crucial Role of the Subpolar North Atlantic for Skillful Decadal Climate Predictions, Geophys. Res. Lett., 51, e2024GL109415, https://doi.org/10.1029/2024GL109415, 2024.

Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., et al.: An evolving Coupled Model Intercomparison Project phase 7 (CMIP7) and Fast Track in support of future climate assessment, Geosci. Model Dev., 18, 6671-6700, https://doi.org/10.5194/gmd-18-6671-2025, 2025.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., et al.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G. et al.: Taking climate model evaluation to the next level, Nat. Clim. Change, 9, 102-110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Gu, Q., Gervais, M., Danabasoglu, G., Kim, W. M., Castruccio, M., et al.: Wide range of possible trajectories of North Atlantic climate in a warming world, Nat. Commun., 15, 4221, https://doi.org/10.1038/s41467-024-48401-2, 2024.

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. G., Domingues, C. M., et al.: Changing State of the Climate System, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC, Cambridge University Press, 287-422, https://doi.org/10.1017/9781009157896.004, 2021.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., et al.: High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6, Geosci. Model Dev., 9, 4185-4208, https://doi.org/10.5194/gmd-9-4185-2016, 2016.

Hausfather, Z.: An assessment of current policy scenarios over the 21st century and the reduced plausibility of high-emissions pathways, Dialogues on Climate Change, 2, 26-32, https://doi.org/10.1177/29768659241304854, 2025.

Hausfather, Z. and Peters, G. P.: Emissions - the "business as usual" story is misleading, Nature, 577, 618-620, https://doi.org/10.1038/d41586-020-00177-3, 2020.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, B. Am. Meteorol. Soc., 90, 1095-1107, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., et al.: The ERA5 global reanalysis, Q. J. Roy. Meteor. Soc., 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Hewitt, H. T., Roberts, M. J., Hyder, P., Graham, T., Rae, J., et al.: The impact of resolving the Rossby radius at mid-latitudes in the ocean: results from a high-resolution version of the Met Office GC2 coupled model, Geosci. Model Dev., 9, 3655-3670, https://doi.org/10.5194/gmd-9-3655-2016, 2016.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Immorlano, F., Eyring, V., le Monnier de Gouville, T., Accarino, G., Elia, D., et al.: Transferring climate change physical knowledge, P. Natl. Acad. Sci. USA, 122, e2413503122, https://doi.org/10.1073/pnas.2413503122, 2025.

IPCC: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., et al., Cambridge University Press, https://doi.org/10.1017/9781009157896, 2021.

Jackson, L. C., Biastoch, A., Buckley, M. W., Desbruyères, D. G., Frajka-Williams, E., et al.: The evolution of the North Atlantic Meridional Overturning Circulation since 1980, Nat. Rev. Earth Environ., 3, 241-254, https://doi.org/10.1038/s43017-022-00263-2, 2022.

King, A. D. and Harrington, L. J.: The inequality of climate change from 1.5 to 2 °C of global warming, Geophys. Res. Lett., 45, 5030-5033, https://doi.org/10.1029/2018GL078430, 2018.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv, https://doi.org/10.48550/arXiv.1412.6980, 2015.

Kosaka, Y. and Xie, S.-P.: Recent global-warming hiatus tied to equatorial Pacific surface cooling, Nature, 501, 403-407, https://doi.org/10.1038/nature12534, 2013.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., et al.: A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, 1909-1918, https://doi.org/10.1002/2016GL072012, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles, in: Adv. Neur. Inf. Process. Syst. (NeurIPS), 30, 6402-6413, 2017.

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., Engelbrecht, F., Fischer, E., Fyfe, J. C., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., and Zhou, T.: Future Global Climate: Scenario-Based Projections and Near-Term Information, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC, Cambridge University Press, 553-672, https://doi.org/10.1017/9781009157896.006, 2021.

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., et al.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth Syst. Dynam., 11, 491-508, https://doi.org/10.5194/esd-11-491-2020, 2020.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Li, C., Wu, J., Wang, Z., Zwiers, F. W., Zhang, X., and Chen, X.: Machine learning helps to strongly reduce future warming uncertainty, Nat. Commun., 17, 3366, https://doi.org/10.1038/s41467-026-70205-9, 2026.

Li, T., Zwiers, F. W., and Zhang, X.: Should we think of observationally constrained multidecade climate projections as predictions? Sci. Adv., 11, eadt6485, https://doi.org/10.1126/sciadv.adt6485, 2025.

Maher, N., Milinski, S., and Ludwig, R.: Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble, Earth Syst. Dynam., 12, 401-418, https://doi.org/10.5194/esd-12-401-2021, 2021.

Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., Delgado-Torres, C., Samsó, M., and Bretonnière, P.-A.: Constraining low-frequency variability in climate projections to predict climate on decadal to multi-decadal timescales – a poor man's initialized prediction system, Earth Syst. Dynam., 13, 1437-1450, https://doi.org/10.5194/esd-13-1437-2022, 2022.

Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., and Ruprich-Robert, Y.: Constraining decadal variability yields skillful projections of near-term climate change, Geophys. Res. Lett., 48, e2021GL094915, https://doi.org/10.1029/2021GL094915, 2021.

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., Taylor, K. E., and Schlund, M.: Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models, Sci. Adv., 6, eaba1981, https://doi.org/10.1126/sciadv.aba1981, 2020.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, Earth Syst. Dynam., 11, 807-834, https://doi.org/10.5194/esd-11-807-2020, 2020.

Michel, S. L. L., Dijkstra, H. A., Guardamagna, F., Jacques-Dumas, V., van Westen, R. M., et al.: Deep learning based reconstructions of the Atlantic meridional overturning circulation confirm twenty-first century decline, Environ. Res. Lett., 20, 064036, https://doi.org/10.1088/1748-9326/add7f0, 2025.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

Michel, S. L. L., Strommen, K., Kröger, J., Roberts, M. J., von-Storch, J.-S. and Christensen, H. Unravelling the role of increased model resolution in simulating surface temperature fields using explainable AI. Under review at AI for the Earth Systems.

Miller, R. L., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Kelley, M., et al.: CMIP6 historical simulations (1850–2014) with GISS-E2.1, J. Adv. Model. Earth Syst., 13, e2019MS002034, https://doi.org/10.1029/2019MS002034, 2021.

Nowack, P. and Watson-Parris, D.: Opinion: Why all emergent constraints are wrong but some are useful – a machine learning perspective, Atmos. Chem. Phys., 25, 2365-2384, https://doi.org/10.5194/acp-25-2365-2025, 2025.

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., et al.: The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6, Geosci. Model Dev., 9, 3461-3482, https://doi.org/10.5194/gmd-9-3461-2016, 2016.

O'Reilly, C. H., Brunner, L., Qasmi, S., Nogherotto, R., Ballinger, A. P., et al.: Assessing observational constraints on future European climate in an out-of-sample framework, npj Clim. Atmos. Sci., 7, 95, https://doi.org/10.1038/s41612-024-00648-8, 2024.

Olonscheck, D., Schurer, A. P., Lücke, L., and Hegerl, G. C.: Large-scale emergence of regional changes in year-to-year temperature variability by the end of the 21st century, Nat. Commun., 12, 7237, https://doi.org/10.1038/s41467-021-27515-x, 2021.

Portmann, V., Chavent, M., and Swingedouw, D.: ClimLoco1.0: CLimate variable confidence Interval of Multivariate Linear Observational COnstraint, Geosci. Model Dev., 18, 9015-9038, https://doi.org/10.5194/gmd-18-9015-2025, 2025.

Portmann, V., Swingedouw, D., Khattab, O., and Chavent, M.: Observational constraints project a ~50% AMOC weakening by the end of this century, Sci. Adv., 12, eadx4298, https://doi.org/10.1126/sciadv.adx4298, 2026.

Pörtner, H.-O., Roberts, D. C., Tignor, M., et al. (Eds.): Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the IPCC, Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009325844, 2022.

Power, S., Lengaigne, M., Capotondi, A., Khodri, M., Vialard, J., et al.: Decadal climate variability in the tropical Pacific: Characteristics, causes, predictability, and prospects, Science, 374, eaay9165, https://doi.org/10.1126/science.aay9165, 2021.

Qasmi, S. and Ribes, A.: Reducing uncertainty in local temperature projections, Sci. Adv., 8, eabo6872, https://doi.org/10.1126/sciadv.abo6872, 2022.

Ribes, A., Boé, J., Qasmi, S., Dubuisson, B., Douville, H., et al.: An updated assessment of past and future warming over France based on a regional observational constraint, Earth Syst. Dynam., 13, 1397-1415, https://doi.org/10.5194/esd-13-1397-2022, 2022.

Ribes, A., Qasmi, S., and Gillett, N. P.: Making climate projections conditional on historical observations, Sci. Adv., 7, eabc0671, https://doi.org/10.1126/sciadv.abc0671, 2021.

Ribes, A., Tessiot, O., Forster, P. M., Gillett, N. P., Masson-Delmotte, V., et al.: Towards annual updating of forced warming to date and constrained climate projections, Nat. Commun., 16, 9214, https://doi.org/10.1038/s41467-025-63026-9, 2025.

Ricard, L., Falasca, F., Runge, J., and Nenes, A.: network-based constraint to evaluate climate sensitivity, Nat. Commun., 15, 6942, https://doi.org/10.1038/s41467-024-50813-z, 2024.

Roberts, M. J., Hewitt, H. T., Hyder, P., Ferreira, D., Josey, S. A., et al.: Impact of ocean resolution on coupled air-sea fluxes and large-scale climate, Geophys. Res. Lett., 43, 10430-10438, https://doi.org/10.1002/2016GL070559, 2016.

Roberts, M. J., Jackson, L. C., Roberts, C. D., Meccia, V., Docquier, D., et al.: Sensitivity of the Atlantic Meridional Overturning Circulation to model resolution in CMIP6 HighResMIP simulations and implications for future changes, J. Adv. Model. Earth Syst., 12, e2019MS002014, https://doi.org/10.1029/2019MS002014, 2020.

Roberts, M. J., Reed, K. A., Bao, Q., Barsugli, J. J., Camargo, S. J., et al.: High-Resolution Model Intercomparison Project phase 2 (HighResMIP2) towards CMIP7, Geosci. Model Dev., 18, 1307-1332, https://doi.org/10.5194/gmd-18-1307-2025, 2025.

Rohde, R. A. and Hausfather, Z.: The Berkeley Earth Land/Ocean Temperature Record, Earth Syst. Sci. Data, 12, 3469-3479, https://doi.org/10.5194/essd-12-3469-2020, 2020.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional networks for biomedical image segmentation, in: MICCAI 2015, Lect. Notes Comput. Sci., 9351, 234-241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T.: A dipole mode in the tropical Indian Ocean, Nature, 401, 360-363, https://doi.org/10.1038/43854, 1999.

Sanderson, B. M., Pendergrass, A. G., Koven, C. D., Brient, F., Booth, B. B. B., et al.: The potential for structural errors in emergent constraints, Earth Syst. Dynam., 12, 899-918, https://doi.org/10.5194/esd-12-899-2021, 2021.

Screen, J. A.: Arctic amplification decreases temperature variance in northern mid- to high-latitudes, Nat. Clim. Change, 4, 577-582, https://doi.org/10.1038/nclimate2268, 2014.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S. M., Wehner, M., and Zhou, B.: Weather and Climate Extreme Events in a Changing Climate, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC, Cambridge University Press, 1513-1766, https://doi.org/10.1017/9781009157896.013, 2021.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., et al.: Uncovering the forced climate response from a single ensemble member using statistical learning, J. Climate, 32, 5677-5699, https://doi.org/10.1175/JCLI-D-18-0882.1, 2019.

Small, R. J., deSzoeke, S. P., Xie, S. P., O'Neil, L. O., Seo, H., et al.: Air-sea interaction over ocean fronts and eddies, Dynam. Atmos. Oceans, 45, 274-319, https://doi.org/10.1016/j.dynatmoce.2008.01.001, 2008.

Sørland, S. L., Schär, C., Lüthi, D., and Kjellström, E.: Bias patterns and climate change signals in GCM-RCM model chains, Environ. Res. Lett., 13, 074017, https://doi.org/10.1088/1748-9326/aacc77, 2018.

Sutton, R. T. and Dong, B.: Atlantic Ocean influence on a shift in European climate in the 1990s, Nat. Geosci., 5, 788-792, https://doi.org/10.1038/ngeo1595, 2012.

Swingedouw, D., Houssais, M.-N., Herbaut, C., Blaizot, A. C., Devilliers, M., et al.: AMOC recent and future trends: A crucial role for oceanic resolution and Greenland melting?, Front. Clim., 4, 838310, https://doi.org/10.3389/fclim.2022.838310, 2022.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., et al.: Past warming trend constrains future warming in CMIP6 models, Sci. Adv., 6, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549, 2020.

https://doi.org/10.5194/egusphere-2026-5397
Preprint. Discussion started: 21 September 2026

UNEP: Emissions Gap Report 2025, United Nations Environment Programme, Nairobi, https://doi.org/10.59117/20.500.11822/47966, 2025.

Venmans, F. and Carr, B.: Literature-informed likelihoods of future emissions and temperatures, Climate Risk Management, 44, 100605, https://doi.org/10.1016/j.crm.2024.100605, 2024.

Wang, H., Qiu, B., Liu, H., and Zhang, Z. Doubling of surface oceanic meridional heat transport by non-symmetry of mesoscale eddies, Nat. Commun., 14, 5460, https://doi.org/10.1038/s41467-023-41294-7, 2023.

Weijer, W., Cheng, W., Garuba, O. A., Hu, A., and Nadiga, B. T.: CMIP6 models predict significant 21st century decline of the Atlantic Meridional Overturning Circulation, Geophys. Res. Lett., 47, e2019GL086075, https://doi.org/10.1029/2019GL086075, 2020.

Yeager, S. G. and Robson, J. I.: Recent progress in understanding and predicting Atlantic decadal climate variability, Curr. Clim. Change Rep., 3, 112-127, https://doi.org/10.1007/s40641-017-0064-z, 2017.

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Danabasoglu, G., Maroon, E., et al.: An outsized role for the Labrador Sea in the multidecadal variability of the Atlantic overturning circulation, Sci. Adv., 7, eabh3592, https://doi.org/10.1126/sciadv.abh3592, 2021.

Zelinka, M. D., Klein, S. A., Qin, Y., and Myers, T. A.: Evaluating climate models' cloud feedbacks against expert judgment, J. Geophys. Res.-Atmos., 127, e2021JD036185, https://doi.org/10.1029/2021JD036185, 2022.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., et al.: Causes of higher climate sensitivity in CMIP6 models, Geophys. Res. Lett., 47, e2019GL085782, https://doi.org/10.1029/2019GL085782, 2020.

Zhang, R. and Sun, C.: A review of the role of the Atlantic Meridional Overturning Circulation in Atlantic Multidecadal Variability and associated climate impacts, Rev. Geophys., 57, 316-375, https://doi.org/10.1029/2019RG000644, 2019.