Synthesis report documenting the new I4C blending strategies

Work Package:	5
Deliverable:	5.2
Due Date:	30.04.2025 (M30)
Submission Date:	30.04.2025 (M30)
Dissemination Level:	Public
Type:	Report
Responsible:	CNRS
Author(s):	Rémy Bonnet (CNRS), Roberto Bilbao (BSC), Pablo Ortega (BSC), Julien Boé (CNRS) Alex Lenkoski (NRS), Rashed Mahmood (DMI), Shuting Yang (DMI), Bo Christiansen (DMI)

Funded by the European Union

1 Summary for Publication

This document provides a description of the blending methods developed in Task 5.2 to produce seamless predictive climate information, following the guidelines provided in Task 5.1. The blending methods presented in this document will be then compared in terms of consistency and usefulness for several case studies in Task 5.3.

2 Contribution to the top-level objectives of Impetus4Change

This deliverable contributes to the following I4C specific objectives:

SO5: Develop seamless homogeneous predictive information from very short (subseasonal) to climate change (several decades) timescales, both at the global and regional scale and additionally i) advance novel approaches to blend and align forecast information across timescales in a way that improves the forecast skill and therefore underpins strategic decision-making and ii) provide this information to I4C demonstrator projects and stakeholder with clear guidance and best practices.
How: By presenting and demonstrating the usefulness of the blending methods developed within WP5.

3 Detailed Report

3.1 Introduction

The recent development of blending methods shows potential benefits for improving our information about the future weather and climate, in timescales that range from a few weeks to several years to decades ahead. Although the data used differ depending on the forecast timescale, the common idea behind these "blending" methods is to combine the multiple forecast sources available to homogenize the information available to users. By leveraging the strengths of each data source, these methods aim to reduce uncertainties and enhance the reliability of forecasts across all lead times, from days to decades.

Blending methods are gaining increasing interest at decadal timescales, as the implementation of adaptation policies requires relevant information about climate evolution over the near-term future (Kushnir et al., 2018). Three sources of information are available to provide relevant climate information over the near-term future. Non-initialized ensemble of climate simulations that provide seamless climate evolution from the historical period to the end of the 21st century but encompassing the full range of uncertainty relative to internal climate variability. Initialized decadal predictions that aim to reduce this uncertainty by initializing the climate model

simulations from an estimate of the observed atmospheric and oceanic state, which intends to phase the simulated and observed climate variability modes and to correct errors in the model's response to forcing. However, their added-value to non-initialized climate projections can be small after a few years (e.g. Yeager et al., 2018) and they are usually limited to 5 to 10 years. They are also subject to initial shocks and drift due to the initialisation (e.g. Sanchez-Gomez et al., 2016). Finally, observations can also be used to constrain the climate evolution over the next decades. Taking advantage of these different sources of information to deliver seamless climate projections — meaning continuous and without artificial shifts in distribution— for the near-term future, with narrow uncertainty related to internal climate variability, remains a challenge.

To address this challenge, several methods have been recently developed to blend information from non-initialized ensembles of climate simulations with the observed climate state or decadal predictions in order to constrain certain aspects of internal climate variability in large ensembles of non-initialized transient climate projections. Befort et al. (2020) and Mahmood et al. (2021) explore this idea by developing methods based on the subselection of non-initialized climate projections from large ensembles based on their agreement with sea surface temperature patterns from initialized decadal predictions. Another method was proposed by Befort et al. (2022) to blend information from decadal prediction and transient historical simulations by concatenating them. This can, however, lead to some inconsistencies affecting the resulting dataset before and after their transition point.

This deliverable describes the blending strategies that have been developed for time-to-event forecasts and decadal timescales in Task 5.2. These results are provided in the following sections, as individual contributions from the different institutions involved in WP5.

3.2 Work Carried Out

New blending method for time-to-event forecasts (NRS)

At NRS, we have studied methods for blending probabilistic time-to-event forecasts, i.e. forecasts that estimate the timing of a specific event. In meteorology, this can, e.g., be a forecast of when we can expect the first frost of the winter season, the onset of the rainy season, a sudden stratospheric warming or a drought period.

Our research (Cunen et al., 2025) explores how to blend time-to-event forecasts from different numerical weather prediction (NWP) systems using a statistical survival model framework. In the field of medicine and finance, survival methods are widely used for modelling onset of events (see e.g. Kalbfleisch and Prentice, 2002). Our study is one of the first exploring the use of survival methods within the field of meteorology. Survival models are particularly suitable for dealing with censored observations, which in the meteorological context occurs when the target event does not happen for an ensemble member before the maximum lead time of the forecast.

We use the timing of the first hard freeze of the winter season for locations in Norway as an example application, and define hard freeze as a day with mean daily temperature below 0 °C. We consider blending of seasonal temperature forecasts from ECMWF, DWD, CMCC, Météo France and UK MET office with lead times around

6 months[1], with ECMWF subseasonal forecasts of temperature with lead time 47 days[2]. The seasonal forecasts are treated as one multi-model ensemble of around 150 members, while the subseasonal forecasts have 11 members.

From each ensemble member, we extract the first occurrence of hard freeze of the season (or the time of censoring) and use survival methods to form one probabilistic distribution for each of the two forecast sources (seasonal and subseasonal) (Figure 1). We then explore different blending methodologies for combining the resulting distributions, both established methods such as Linear Pooling and Beta Pooling (Gneiting and Ranjan, 2013; Baran and Lerch, 2018), and methods of our own making, Gaussian Pooling and Hazard Blending (Cunen et al., 2025). Model parameters were estimated from observation-hindcast pairs.

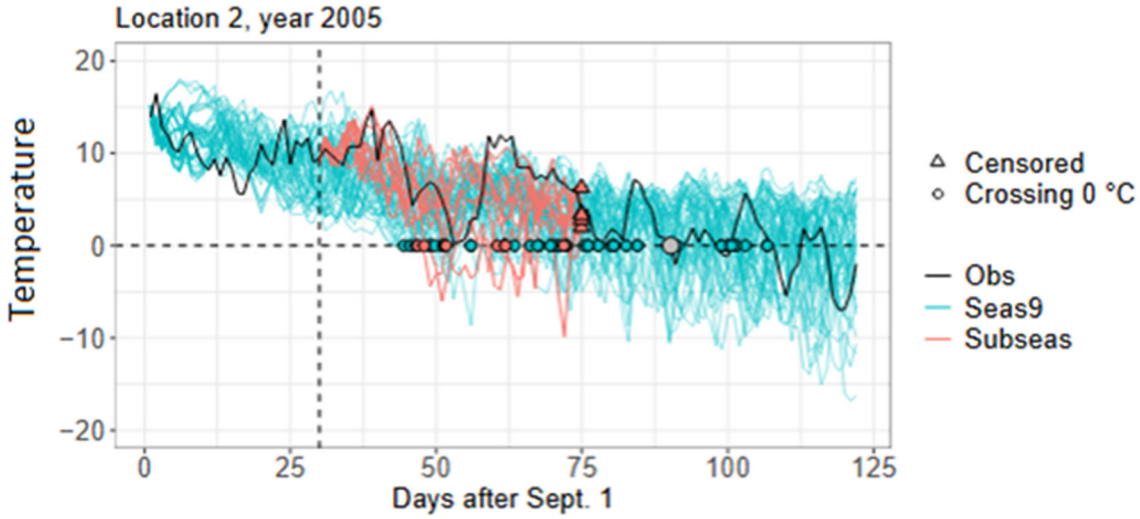


Figure 1. Temperature trajectories from seasonal forecasts (blue) and subseasonal forecasts (red) with two different issue dates (day 1 and day 29) for a selected study location. The observations (black) are derived from Lussana, (2020). The first crossing of 0 °C is extracted for each ensemble member (circles), or the time of censoring if the ensemble member never crosses 0 °C (triangles). The timing data are used to form two probabilistic time-to-hard-freeze distributions that are combined by using different blending strategies.

To evaluate the effectiveness of the blending methods, we conducted a comprehensive simulation study. This was done by creating simulated temperature trajectories mimicking properties of the seasonal and subseasonal temperature forecasts. The simulation study allowed us to test the blending methods under controlled conditions, with varying levels of forecast bias, dispersion and number of hindcast-observation pairs. In addition, we tested the methods for a real-world case study, with temperature data from Norway and Fennoscandia.

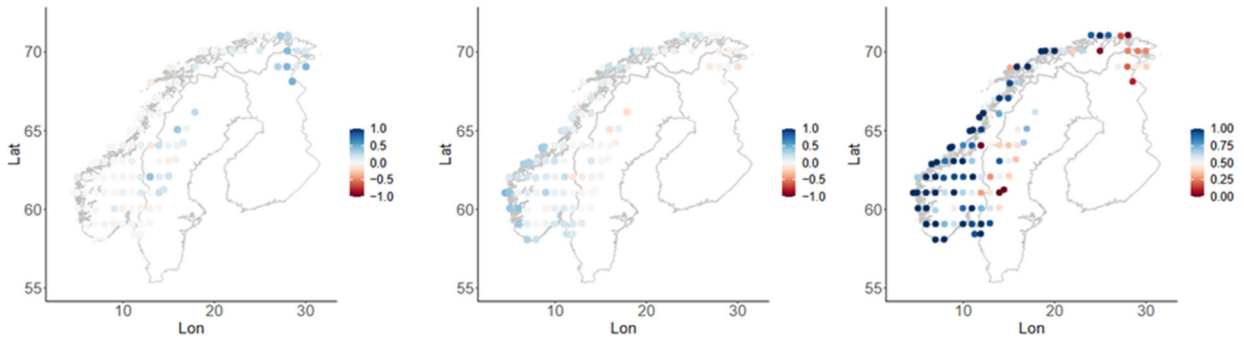
The simulation study demonstrated that combination forecasts generally improve upon single-source forecasts when the sources are balanced in terms of noise-to-signal ratio and there is enough hindcast data. However, when the number of hindcast-observation pairs is low or the forecasts are highly unbalanced, all blending

[1] C3S Seasonal Forecasts: dataset documentation - Copernicus Knowledge Base - ECMWF Confluence Wiki

[2] https://confluence.ecmwf.int/display/S2S/ECMWF+model+descript

methods struggle, with more complex approaches performing worse than simpler alternatives.

For the real case study from Fennoscandia, most of the blending approaches showed skill compared to single-source forecasts and climatology for the time-to-hard-freeze forecasts. Simpler methods, such as Linear Pooling, outperformed more advanced techniques, likely due to the limited number of hindcast-observation pairs. In Figure 2, the skill scores of the Linear Pooling method is shown for the case study. Particularly at locations far from the coast, and in Northern Norway, the blending method has skill compared to the single-source forecasts. These are locations where hard freeze typically occurs closer to our selected issue date, October 1.



(a) IBS skill score relative to the seasonal forecast (and climatology).

(b) IBS skill score relative to the subseasonal forecast.

(c) Weight ω on the seasonal forecast in the LP (ML) method.

Figure 2. The skill of the linear pooling (LP) blending method relative to the seasonal forecast (a) and subseasonal forecast (b) for time-to-hard-freeze predictions in Fennoscandia, with issue date October 1. Here, 1 is a perfect score, while scores below 0 means that the single-source reference forecast performs better. Figure (c) shows the weight that is given to each of the two forecast sources during blending, with blue indicating high weight on the seasonal forecast and red indicating high weight on the subseasonal forecast.

While the time-to-hard-freeze application was used to motivate our research, the blending methods and simulation framework can be used to study other time-to-event forecasts in meteorology. The methods and results from our work are reported in a paper that is currently in review in a scientific journal. A preprint is available on arxiv (Cunen et al., 2025).

New blending method to predict winter temperature over Europe and Northern Asia on multi-annual to decadal timescales (DMI)

Predicting the winter climate over Europe and northern Asia remains challenging especially on multi-annual to decadal timescales. The initialized climate predictions have shown limited success in predicting the winter climate variability over Eurasia as the skill of these predictions vanes quickly after the first forecast year. Eurasia is also the region where internal climate variability is strongly modulated by the North Atlantic Oscillations (NAO) (e.g. Trigo et al., 2002; Ye et al., 2022). We explore here whether this observed teleconnection between NAO and the surface air temperature (SAT) can be used for constraining climate model simulations in order to provide skillful estimates of winter climate in this region for the next ten years.

For this, at DMI we have developed a new methodology for constraining variability in historical simulation ensembles by exploiting the observed teleconnection between NAO and surface air temperature. This new methodology is complementary to the other methodologies in which sea surface temperature anomalies were used as a constraint (e.g. Mahmood et al., 2022; Donat et al., 2024). The new methodology presented here involves computing regression between winter season NAO and surface air temperature during a 20 year window (Figure 3). The observed and model member simulated spatial distribution of the regression patterns are compared in order to rank the model members. The highest ranking members are then used to make predictions after the constraining period. This procedure is repeated yearly, for example, to predict temperatures from 1971 and onwards the constraining period used is 1951 to 1970 and similarly to predict temperature from 1972 and onwards the constraining period is 1952 to 1971 and so on. Based on this procedure, we build a hindcast of 40 initializations for evaluating the skill of the predictions.

The constrained ensemble based on this methodology can make skillful predictions of seasonal to multi-annual mean winter climate over Europe and northern Asia. The constraint evaluates simulated and observed NAO-temperature teleconnection patterns during 20 year windows prior to making predictions. The resulting top ranking 10 members are used to make predictions for the next 10 years. We find here that the constrained ensemble is skillful in predicting winter SAT on multi-annual to decadal timescales over Euroasia (Figure 4). We also find that the constrained multi-annual predictions yield higher correlation skill compared to both the unconstrained ensemble and the initialized predictions, especially up to 5 year mean winter SAT predictions. Therefore, we argue that for the Eurasian region, the constraint could provide skillful winter SAT predictions on multi-annual timescales with minimal cost compared to the state of the art initialized decadal climate predictions for which added value from initialization wanes quickly after the first forecast year.

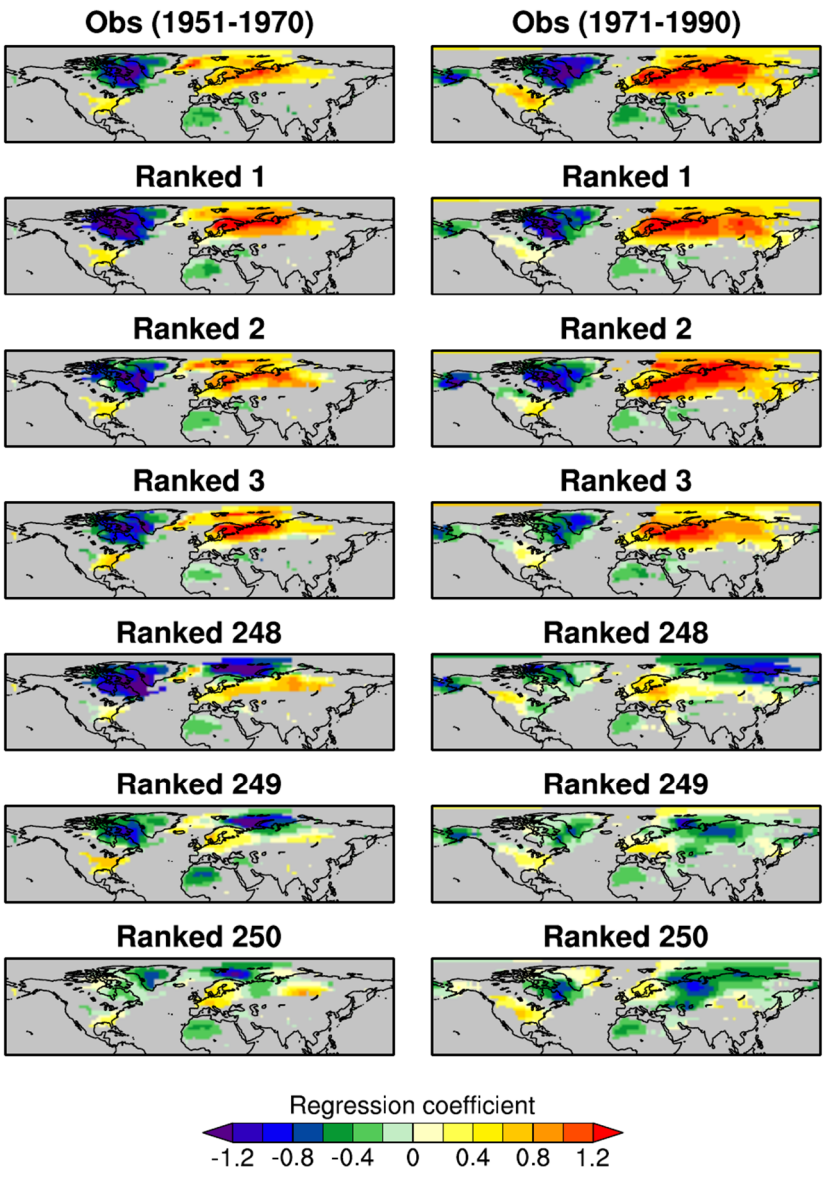


Figure 3. A comparison of the regression pattern between winter time (Dec-Jan-Feb) NAO and the surface air temperature during the 20 year period. The top rows show the observed relationship between NAO and temperature (using ERA5; Hersbach et al., 2020) while the subsequent rows show the relationship in individual model members (a total 250 members were used). For each start-date a spatial pattern correlation is calculated between the observed and simulated NAO-temperature regressions to sort the members from highest to lowest ranking members (shown here as top and bottom ranking three members). The top-ranking members can be selected to make predictions after the constraining period. Here we only show comparisons for two start dates, however, a total of 40 start-dates (i.e. 1971 - 2010) were used in building a full hindcast for evaluation.

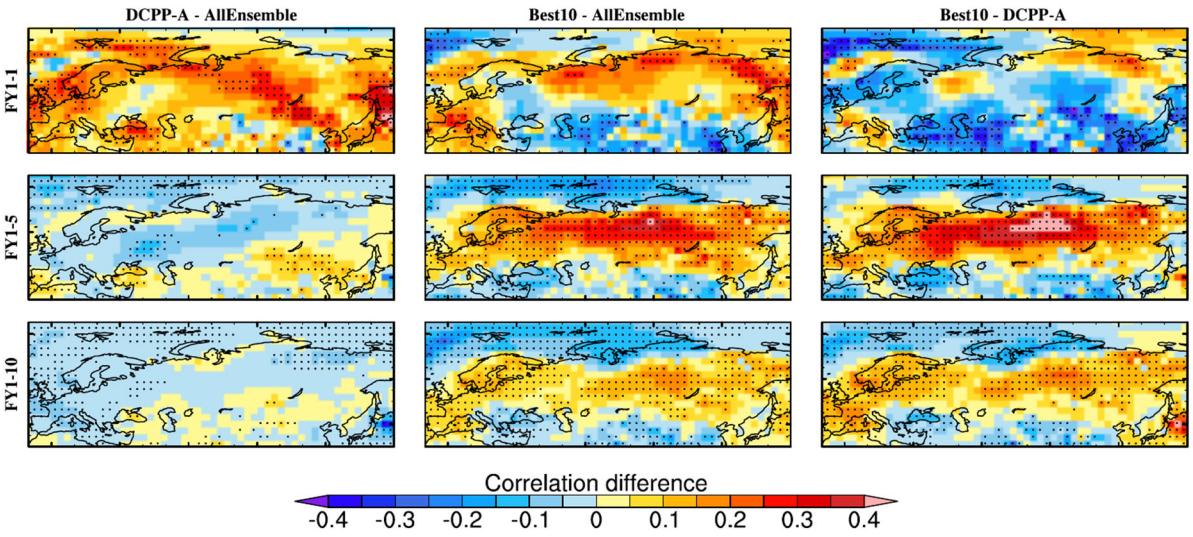


Figure 4. ACC difference for winter surface air temperature between DCPP-A and historical ensemble (left panels), Best10 and historical ensemble (center panels) and Best10 and DCPP-A (right panels) for different forecast periods. Stippling indicates regions where the correlation difference is statistically significant at the 95% confidence level.

New blending method to provide seamless climate information over the near-term future: a case study with winter Mediterranean temperature (CNRS-Cerfacs)

CNRS-Cerfacs has developed a novel blending methodology to explore the potential benefits of combining the three available sources of information —observations, initialized decadal predictions, and non-initialized climate projections— to provide relevant information on near-term climate change with reduced uncertainty related to internal climate variability. In this report, we present a case study for winter surface temperature over the Mediterranean region.

We used an ensemble of 92 members from 6 prediction systems (DCPP-A; Boer et al 2016) and 163 members from the equivalent models for transient historical simulations (Eyring et al., 2016), summarized in Table 1. The blending method developed in this study is based on two steps (described below).

Model	Number of simulations (**historical** / *prediction system*)
CNRM-CM6-1(historical) CNRM-ESM2-1(prediction system)	**30** / *25*
EC-Earth3	**15** / *15*
MIROC6	**50** / *10*
MRI-ESM2-0	**12** / *12*
NorCPM1	**30** / *20*
IPSL-CM6A-LR	**26** / *10*

Table 1. Summary of the models used from the 6 predictions system and the corresponding model for the transient historical simulations. Only the data from the CNRM-Cerfacs modelling center do not come from the same version of the model for the prediction system and the historical simulations.

The first step is to select from the historical simulations those that are the closest to observed climate indices before the start of the forecast. Four climate indices are tested for this first selection (step 1):

- **The Atlantic Multidecadal Variability (AMV) index**, which described the evolution of the leading mode of multidecadal variability in the North Atlantic Ocean (Kerr, 2000). The AMV has been linked to many observed low-frequency regional climate variations, including European precipitation and temperature (Sutton & Dong, 2012). The AMV index is defined as the average SST over the North Atlantic (0–60° N, 80°W–0° E) after the removal of the externally forced signal (Trenberth and Shea, 2006). A low-pass filter is then used to retain only the low-frequency variations.

- **The North Atlantic subpolar gyre (SPG) index**, which is a key part of the decadal variability of North Atlantic SST and has been linked to the European climate (e.g. Hermanson et al., 2014). In this study, we define the NASPG index as the average SST over the 15°W–40°W, 50°N–60°N region.

- **The winter (December to February) North Atlantic Oscillation (NAO)**, which is the dominant mode of atmospheric circulation variability in the North Atlantic sector. Winter NAO exerts a strong influence on European weather and climate (e.g. Hurrell et al., 2003). The index is defined as the difference in area-averaged mean sea level pressure (MSLP) between a southern box (20–55° N, 90°W–60° E) and a northern box (55–90° N, 90°W–60° E) in the North Atlantic (Stephenson et al., 2006; Baker et al., 2018).

- **The 9-year average global sea surface temperature (GSST) pattern**. This index has been proposed in previous studies to constrain low-frequency internal climate variability in surface temperature by selecting the simulations closest to

the observed spatial distribution of sea surface temperature based on spatial correlation (e.g. Befort et al. 2020; Mahmood et al. 2022).

The indices based on SST are evaluated against the NOAA Extended Reconstructed SST V5 (ERSSTv5; Huang et al., 2017) observed dataset of sea surface temperature. The winter NAO index is evaluated against the ERA5 reanalysis (Hersbach et al. 2020).

The selection method consists of two steps. First, we select a subset of historical simulations that are closest to an observed climate index relevant to the variable and region of interest over the Y years preceding the forecast. In the example shown in Figure 5, we use the Atlantic Multidecadal Variability (AMV) as the relevant climate index and set Y=20 years. We select a first subset of size N_1=30, which is used in step 2, and a second subset of size N_2=20, which serves as a forecast based solely on the selection of historical simulations from observations (referred to as $Subset_{OBS}(N_2)$ in Figure CNRS_1). The forecast $Subset_{OBS}(N_2)$ shows a reduced spread in the distribution of winter surface temperature over the Mediterranean region compared to the full historical ensemble, and its spread is relatively close to that of the decadal predictions (Figure 1, step 1). Then, the second step consists of refining the selection of N_1 historical simulations based on observations by selecting a sub-ensemble of N_2 simulations (Figure CNRS_1; step 2). This is done by selecting the N_2 simulations from the $Subset_{OBS}(N_1)$ that are the closest to the winter surface temperature of the ensemble mean of the hindcasts over the region of interest, here the Mediterranean region. The resulting forecast, $Subset_{OBS+Hindcast}(N_2)$, derived from these two steps, further reduces uncertainty compared to $Subset_{OBS}(N_2)$ alone, as illustrated in Figure 5.

The historical simulations and the hindcast simulations are compared against $subset_{AMV}$, $subset_{SPG}$, $subset_{NAO}$, $subset_{GSST}$ derived from the first step of the method and $subset_{AMV+hindcast}$, $subset_{SPG+hindcast}$, $subset_{NAO+hindcast}$, $subset_{GSST+hindcast}$ derived from the full method. We also developed $subset_{TAS}$, which derives from the first step of the method and is based on winter surface temperature over the region of interest (MED), to assess whether it is sufficient to only use the variable we are trying to predict to constrain historical simulations. Finally, we also derive $subset_{hindcast}$ that is based on the selection of historical simulations closest to the surface temperature of the hindcasts ensemble mean over the region of interest.

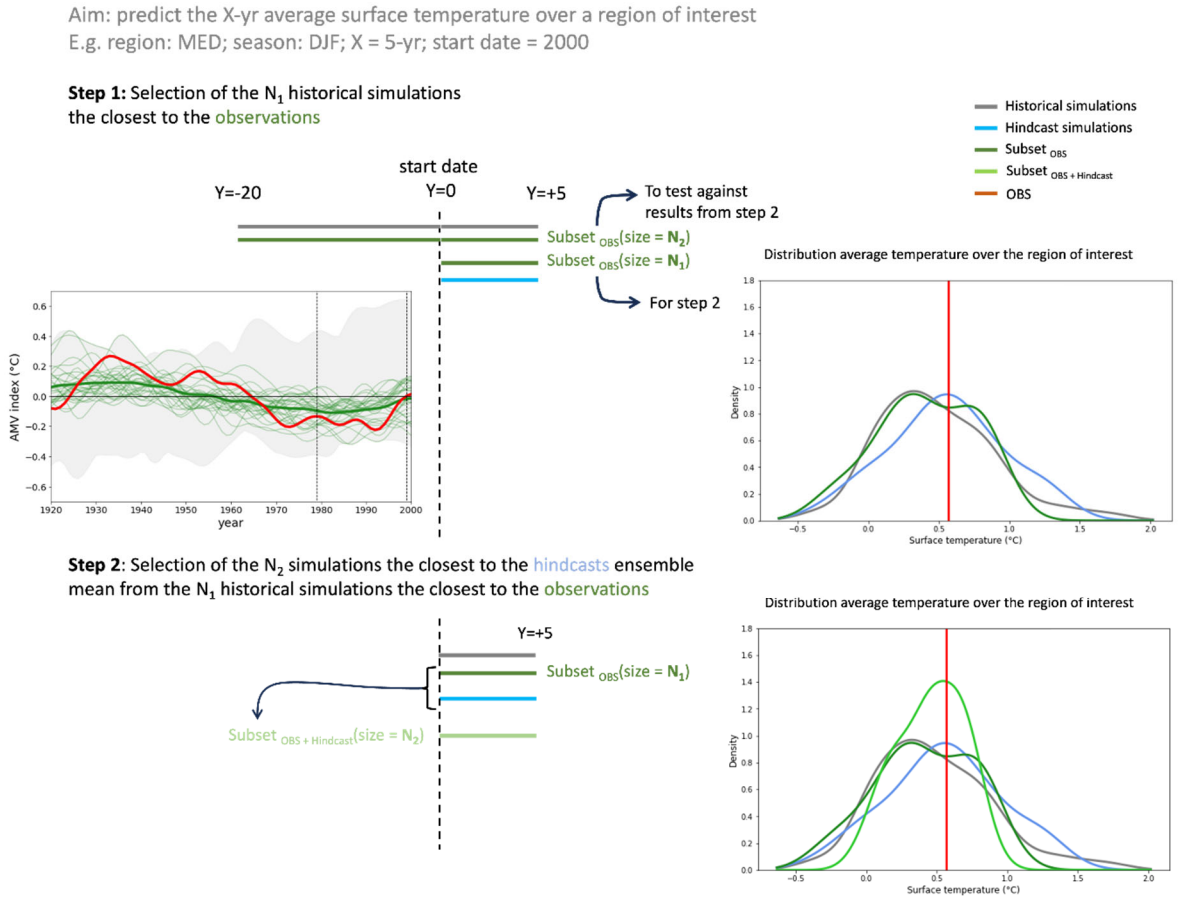


Figure 5. Diagram illustrating the concept and main steps of the blending method. In step 1, the Atlantic Multidecadal Variability (AMV) index from the historical simulations (minimum and maximum in gray) is compared to the ERSSTv5 observational dataset (Huang et al., 2017; green line), with the selection of the best 30 members in green. The histogram of surface temperature predictions are evaluated against the ERA5 reanalysis (Hersbach et al., 2020; red line).

There is an overall important reduction in the spread of average winter surface temperature predictions over the Mediterranean region during the testing period (1967–2000) for all subsets derived from our new blending method, compared to the historical ensemble (Figure 6a). $Subset_{hindcast}$ shows the largest reduction in spread for 5-year prediction, likely because the large number of historical simulations increases the chance of selecting simulations close to the hindcast mean, although this effect is smaller for 10- and 15-year predictions compared to the other subsets. $Subset_{AMV+hindcast}$ and $subset_{SPG+hindcast}$ also show the larger decrease of the spread for the three time horizons in comparison to the other subset derived from the blending method.

Interestingly, the spread of the hindcast simulations is comparable or even larger than the one of the historical ensemble. This suggests that the initialization does not allow to reduce the uncertainty in winter surface temperature due to internal climate variability over this region.

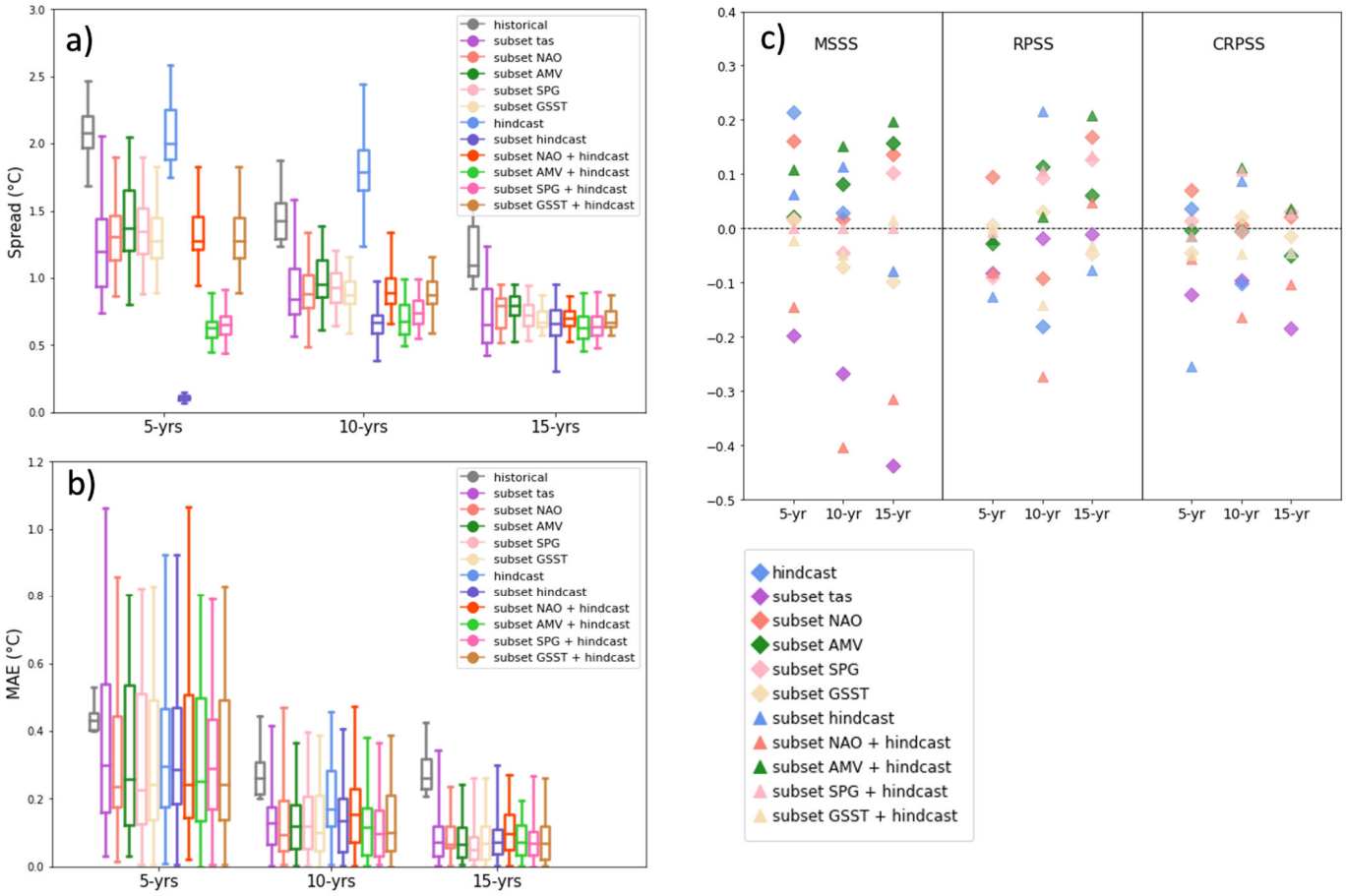


Figure 6. Boxplots of the distribution of the (a) spread and (b) Mean Average Error (MAE) of 5-yr, 10-yr and 15-yr forecast of average winter surface temperature calculated over the Mediterranean region as defined in the IPCC (Iturbide et al., 2020). The spread is defined as the difference between the minimum and the maximum and the MAE is calculated between the observed surface temperature from ERA5 (Hersbach et al. 2020) and the ensemble mean of the different dataset tested. The boxplots are defined with the minimum, 25th percentile, median, 75th percentile and maximum. The spread and the MAE are calculated for each year of the testing period (1967-2000). (c) Mean squared skill score (MSSS), Ranked Probability Skill Score (RPSS) and Continuous Ranked Probability Skill Score (CRPSS) calculated over the 1967-2000 period, using the ensemble mean for the MSSS and the whole ensembles for the RPSS and CRPSS. For these three scores, a positive means a better prediction than the historical ensemble and a negative value a worse prediction.

In addition to this reduction of the spread within the subset ensembles, there is also an overall decrease in the median of the mean absolute error (MAE) of the subsets ensemble mean in comparison to the historical ensemble mean calculated over the testing period (Figure 6b). This reduction in the MAE is particularly important at 10-year and 15 year predictions. $\text{Subset}_{\text{AMV+hindcast}}$ and $\text{subset}_{\text{SPG+hindcast}}$ that show an important decrease in the spread over the testing period also show among the largest MAE decreases between the subsets, highlighting the potential of the method and the predictor used. Although a lower MAE does not imply that observations always fall within the ensemble spread, the overall MAE reduction—particularly for 10- and 15-year forecasts—indicates that the blending method improves the ensemble mean prediction compared to the historical ensemble mean, likely by partially capturing

internal decadal variability beyond external forcing alone. Interestingly, the hindcast ensemble shows an overall decrease in the MAE in comparison to the historical ensemble, which indicates that the initialization allows part of the internal variability to be captured. This also supports the second step of the methodology developed here, i.e. to refine the first selection, based on observations, using the hindcasts ensemble mean.

We also evaluate the prediction of 5, 10 and 15 years average winter temperature over the 1967-2000 period in comparison to the historical ensemble of simulations for three scores: Mean squared skill score (MSSS), Ranked Probability Skill Score (RPSS) and Continuous Ranked Probability Skill Score (CRPSS) (Figure 6c). The MSSS is a deterministic score that informs if the ensemble mean is close to observations, while the RPSS and CRPSS are probabilistic, including how well the forecast captures uncertainty. The results are contrasted, depending on the score. The hindcast ensemble shows a better MSSS than the historical ensemble, but with lower RPSS and small or lower CPRSS, for 5 and 10 years prediction, which seems consistent with the decrease in the MAE (Fig 6b) but the overall larger spread than the historical simulation (Fig 6a). The $subset_{AMV+hindcast}$ shows a better prediction than the historical ensemble based on the MSSS for 5, 10 and 15 years prediction, as well as based on the RPSS and CRPSS for 10 and 15-years prediction. $Subset_{hindcast}$ shows a strong improvement of the 10 year prediction in comparison of the historical ensemble. Interestingly, the $subset_{NAO+hindcast}$ shows a deterioration in the forecast in comparison to the historical ensemble whereas the $subset_{NAO}$ shows an overall improvement of the forecast.

Finally, we evaluate the residual ACC for the hindcast, the $subset_{AMV}$ and the $subset_{hindcast}$ and $subset_{AMV+hindcast}$ as they appear to be the best subset based on the previous evaluation (Figure 7). The residual ACC is calculated following Smith et al. (2019) and measures how well the subset ensembles capture the observed internal variability that is not captured by the ensemble mean of the historical ensemble. The hindcast shows no significant added value for the 10-year forecast, except in parts of northern Europe and the southwest of Spain. By selecting the historical simulations closest to the hindcast ensemble mean of the average winter temperature over the Mediterranean region, a part of the western Maghreb region shows positive significant residual correlation in $subset_{hindcast}$. $Subset_{AMV}$ shows significant positive residual correlations over Spain and a large part of France, highlighting the added value of constraining historical simulations using this climate mode of variability to predict winter temperatures in these regions. By adding a second selection based on the hindcast, $subset_{AMV+hindcast}$ shows higher correlation over Spain as well as a part of the North Maghreb region.

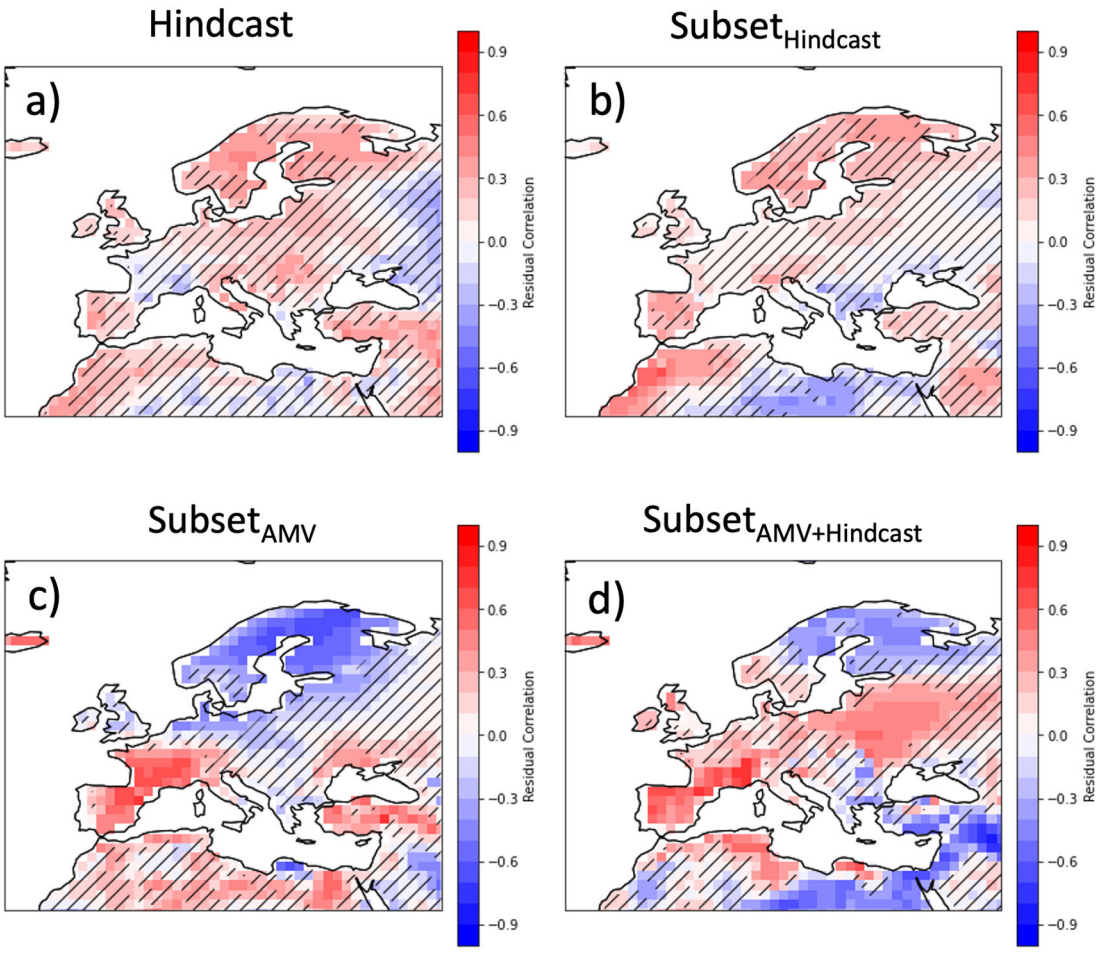


Figure 7. Residual Anomaly Correlation Coefficient (ACC; Smith et al. 2019) calculated from the ensemble mean of 10-year prediction of winter temperature for the (a) hindcast ensemble, (b) $Subset_{hindcast}$, (c) $subset_{AMV}$ and (d) $subset_{AMV+hindcast}$ over the 1967-2000 period. The hatched regions indicate statistically non-significant values ($p < 0.05$) using a permutation test.

The new methodology proposed here appears to be a promising approach for providing seamless and relevant climate information over a region of interest for the near-term future, with reduced uncertainty associated with internal climate variability, while avoiding the drift caused by the shock of the initialization in decadal prediction. One strength of this method is that it can be easily applied to other regions or variables of interest. For this case study, it appears that the $subset_{AMV+hindcast}$, $subset_{SPG+hindcast}$ and $subset_{hindcast}$ are the most effective for predicting winter surface temperature over the Mediterranean region for 5, 10 and 15 year predictions. The added value for Spain highlighted in Figure 7c and 7d is of particular interest for the I4C project, as Barcelona is one of the demonstrator cities.

New blending methodology to provide seamless climate information for the next 30 years (BSC)

The BSC has developed a novel blending methodology that combines information from decadal predictions and climate projections to provide seamless climate

information for the next 30 years. This methodology is motivated by the fact that, during the initial decades of climate projections - particularly at regional scales - the dominant source of uncertainty is the internal variability of the climate system (e.g., Hawkins and Sutton, 2009). Since decadal predictions are initialised from the observed state, by aligning the model's climate variability with real-world observations, they tend to be more skillful than climate projections. However, climate predictions typically extend only 10 years into the future. To provide the best possible information for the coming decades, it is therefore essential to efficiently combine decadal predictions with climate projections. As noted by Befort et al. (2022), simply concatenating decadal climate predictions and climate projections is not straightforward and may lead to inconsistencies during the transition period. Furthermore, as demonstrated in Deliverable 5.1, differences exist in both the statistical properties and physical processes of sea surface temperature (SST) variability at the transition between decadal predictions and climate projections that need to be tackled for an effective blending.

BSC's blending approach builds on the analog-based constraining method developed by Mahmood et al. (2021), which has been shown to effectively improve the predictive capacity of climate projections by reducing the uncertainties related to internal climate variability. However, an important difference is that while in Mahmood et al. (2021) only the ensemble mean is considered, our new methodology blends the whole ensemble in the projections and predictions, which makes it suitable for probabilistic predictions, including of climate extremes. To implement this and address the inconsistencies between predictions and projections additional considerations are required. While the method in Mahmood et al. (2021) ensures physical consistency by ranking climate projection members based on their similarity to the multi-model ensemble mean of decadal predictions and selecting the highest-ranked members, our approach applies this ranking individually to each member of the decadal prediction ensemble. Thus, by selecting the projection in closest agreement with each of the individual forecasts, we generate an ensemble of climate projections that matches the size of the decadal prediction ensemble. This allows us to also preserve the statistical properties of both ensembles.

Several choices must be made when determining which climate projection members best align with the variability in decadal predictions. The sensitivity to some of these choices has been investigated:

- **Selection method:** We have used *spatial pattern correlations* (*centered* and *uncentered*) and *Euclidean distance*.
- **Region where SST anomaly patterns are derived:** We seek for analogs based on *global SST anomaly fields*, which has been previously shown to provide the best results at the global scale, and *North Atlantic SST anomaly fields* (0°–60°N and 80°W-0), which has the potential to improve the prediction skill over Europe, the key region of interest for I4C.
- **Forecast range and temporal aggregation for computing the anomalies:** Because the objective is to blend or stitch predictions and projections, we focus on forecast year ten, the last year available in the predictions. The anomalies have been computed over two different aggregation periods: the *average of the first nine years* and the *average of years five to nine*.

- **Repetition in model ensemble selection:** We have tested two selection criteria, *allowing repetition* of members in the selected ensemble of projections, and *not allowing repetitions*.

Model	DCPP members	Historical Members
CanESM5	10	25 p1 + 25 p2
CMCC-CM2-SR5	10	10
EC-Earth3 i4	10	22
IPSL-CM6A-LR	10	33
MIROC6	10	22
MPI-ESM1-2-HR	10	10
NorCPM1 p1	10	30
NorCPM1 p2	10	-
Total	80	174

Table 1. Climate models and ensemble members considered in this deliverable.

To assess the sensitivity of these choices and the effectiveness of this blending methodology, we apply it to historical simulations with decadal hindcasts for start dates covering the period 1960–2014, during which predictive skill can be evaluated against observations. We use an ensemble of 80 members from eight decadal prediction systems and 177 members from the equivalent models for historical simulations (Table 1). This decision of only considering simulations from the same models in both ensembles seeks to maximise the physical consistency between the ensembles. Anomalies are computed relative to the 1980–2010 period. For decadal predictions, the lead-time-dependent climatology is removed to correct for drift when computing the anomalies.

A first consideration of this method is that for each decadal ensemble member we determine the most similar historical member allowing for any model member to be chosen (not just from the same model). This means that for each of the 80 decadal prediction members, the 177 historical members are ranked by similarity for each start date. Then by selecting the highest ranked members an ensemble of 80 historical members is formed per start date. Figures 8 a and b show the distribution of the highest ranked members for the centered pattern correlation and Euclidean distance methods based on the Global Ocean pattern for each start date. While for the centered pattern correlation the distribution is consistent for each forecast time, for the Euclidean distance that the distribution varies with start date, and generally worse members tend to be selected in the first half of the start dates. This happens as a result of the climatological reference period chosen, as anomalies center around zero as they approach the central years of the reference period. An alternative metric to the

Euclidean distance that could circumvent this issue would be the use of centered root mean square difference, which we will test in future analyses. The uncentered pattern correlation shows a similar behaviour than the Euclidean distance, as discussed in Donat et al. (2024), which recommend the centered approach. We find that in general the highest ranked historical members for each decadal prediction member are from the same model approximately 50% of the time for either method, thus ensuring a high level of physical consistency (figure 8 c and d). It is also important to check if for some start dates the same historical member is selected as the highest ranked member for multiple decadal prediction members (figure 8 c and d). This could pose a problem if the same members are selected recurrently, which could strongly reduce the ensemble spread. Encouragingly, we find that approximately 50-60% of the highest ranked historical ensembles are formed by different members. Similar results are found using these methods based on the North Atlantic Ocean pattern (not shown).

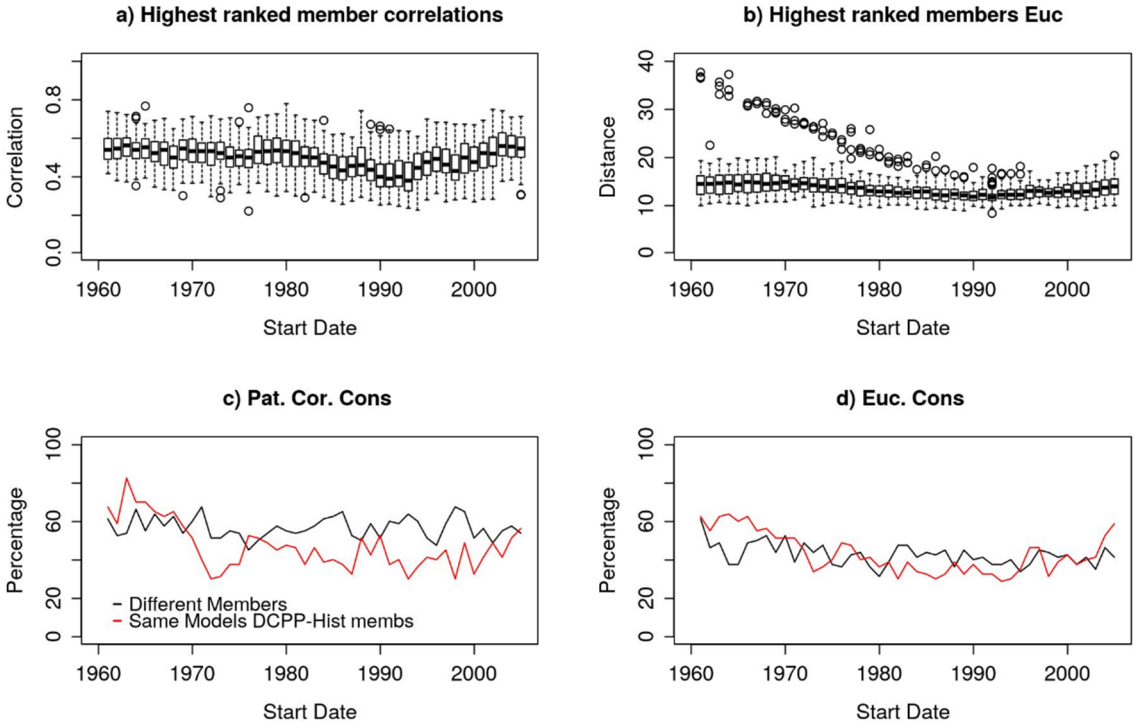


Figure 8. Member selection metrics for using the centered pattern correlation and Euclidean distance methods based on the Global Ocean pattern. a) and b) show the distribution of highest ranked correlations and lowest Euclidean distance respectively, for each start date. c) and d) show the percentage of members selected that are different and the percentage of historical members that select the same model as the decadal predictions.

Figure 9 shows the temporal mean of the intra-ensemble standard deviation computed for the SST anomaly fields in the decadal predictions forecast year 10, the full and the constrained historical ensemble, this later based on both the Global and North Atlantic regions and using both the centered pattern correlation and the Euclidean distance methods. The patterns of the intra-ensemble standard deviation are very similar for the decadal predictions forecast year 10 and the full historical ensemble, with the exception of the North Atlantic region that shows substantially larger values in the historical simulations than in the decadal predictions. This is most

likely because the North Atlantic is a region where the impact of initialisation persists on decadal times-scales (e.g. Meehl et al., 2014 and references therein). We find that constraining the historical ensemble based on the decadal predictions reduces the intra-ensemble spread in the North Atlantic, making it more similar to the one of the decadal predictions. Regarding the methodological choices, we find that constraining based on the Global or the North Atlantic SST patterns has a small impact while the metric used to determine the analogs largely changes the results, with the Euclidean distance yielding a much smaller intra-ensemble spread than the centered pattern correlation approach. We also tested the *uncentered pattern correlation* approach (not shown) which gave similar results, however, as shown by Donat et al. (2024), this approach may lead to unrealistic results as it is strongly sensitive to the long-term trend, and therefore the centered approach is recommended instead.

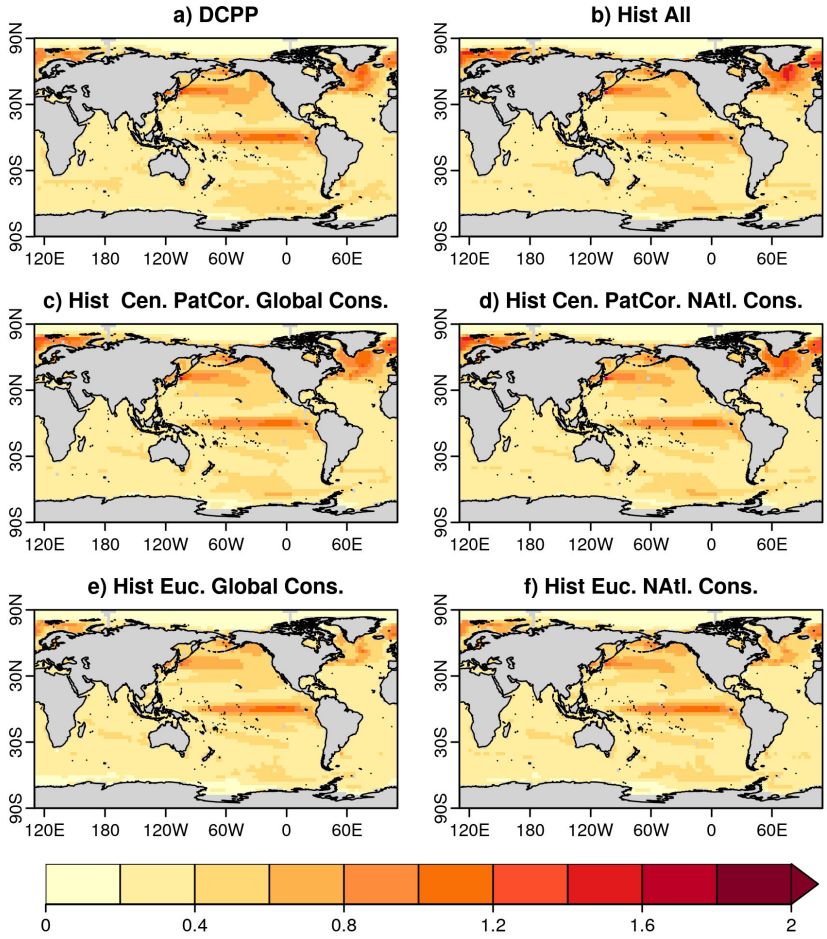


Figure 9. Temporal mean of the multi-model intra-ensemble standard deviation for the SST anomaly fields (°C) in (a) the decadal predictions at forecast year 10, (b) historical simulations, and (c-f) four constrained historical ensembles with different choices. The period covered by the experiments is 1970-2014.

Figure 10 shows the global mean sea surface temperature for the decadal predictions at forecast year 10 and for the full and the analog-based constrained historical ensemble using the same four methodological choices used in Figure 9. This reflects the agreement between the ensembles at the forecast time that the blending is implemented. The ensemble spreads (herein defined as the minimum and maximum values within the ensemble) show small differences between the full historical and

prediction ensembles, which is expected because most of global-scale variability is of externally forced origin. Despite this, all analog-based methods achieve some spread reduction, bringing the ensemble spread closer to that of the decadal prediction ensemble. This indicates that the analog-based blending method improves the statistical consistency between the predictions and historical simulations at their transition point. While constraining based on the global SST patterns or the North Atlantic alone does not have a clear impact, using the Euclidean distance method reduces more the ensemble spread in comparison to the uncentered pattern correction method, leading to a small over reduction of the ensemble spread with respect to the decadal predictions.

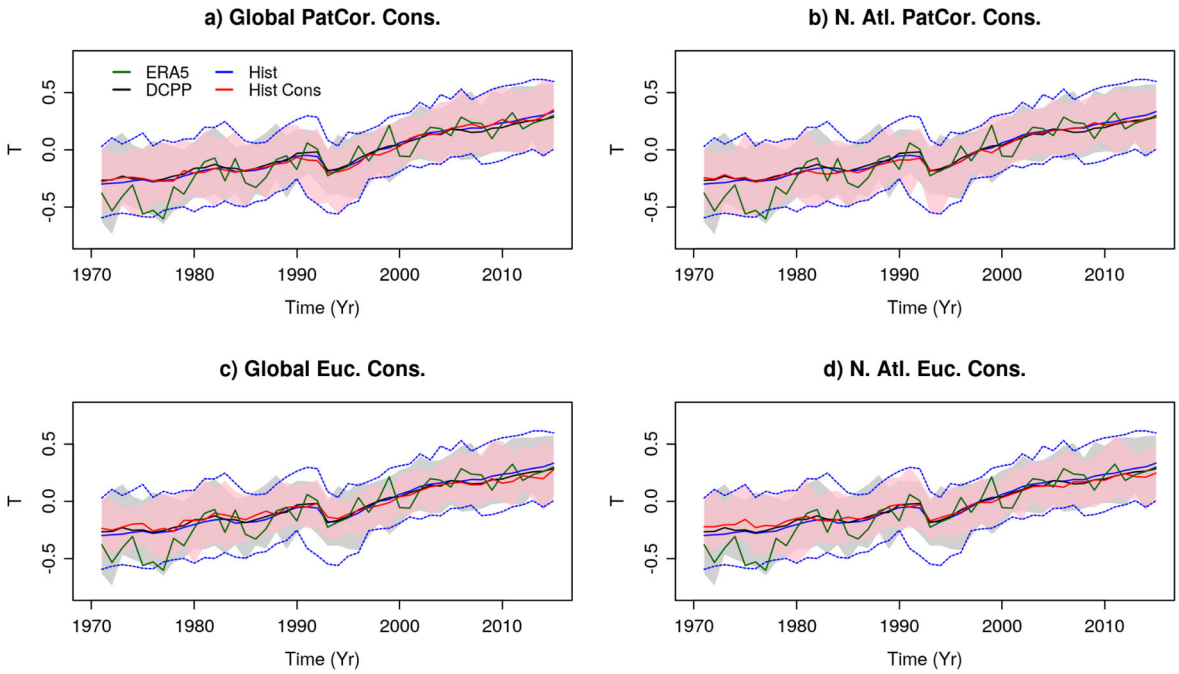


Figure 10. Global mean SST (°C) at the transition between the decadal predictions and historical simulations for the full and constrained ensembles using four methods. Gray and pink shading are the minimum-maximum range of the decadal prediction and the historical constrained ensembles, respectively. Dashed blue lines are the minimum-maximum range of the full historical ensemble.

Figure 11 is similar to Figure 10 but for the North Atlantic mean SSTs, and nicely illustrates that in that region the analog-based method is much more effective in constraining the spread of the full historical ensemble to match the one of the predictions. While we find that in this region all the subselection methods reduce the historical ensemble spread, bringing them closer to the spread of the decadal predictions and thus improving the statistical consistency of the blended product, there are again important differences associated with the metric considered. Using the Euclidean distance method leads to a slight over reduction of the ensemble spread with respect to the decadal prediction ensemble for all start dates. By contrast the centred pattern correlation approach tends to overestimate the spread, in particular, for the start dates prior to year 2000, which misrepresent the lower bound. For start dates after 2000 the centred pattern correlation approach works more effectively in constraining the historical ensemble. This could be related to the transition from negative to positive anomalies in the Subpolar North Atlantic. Again, considering the constraining based

on the global ocean or the North Atlantic does not seem to have a clear effect for this index.

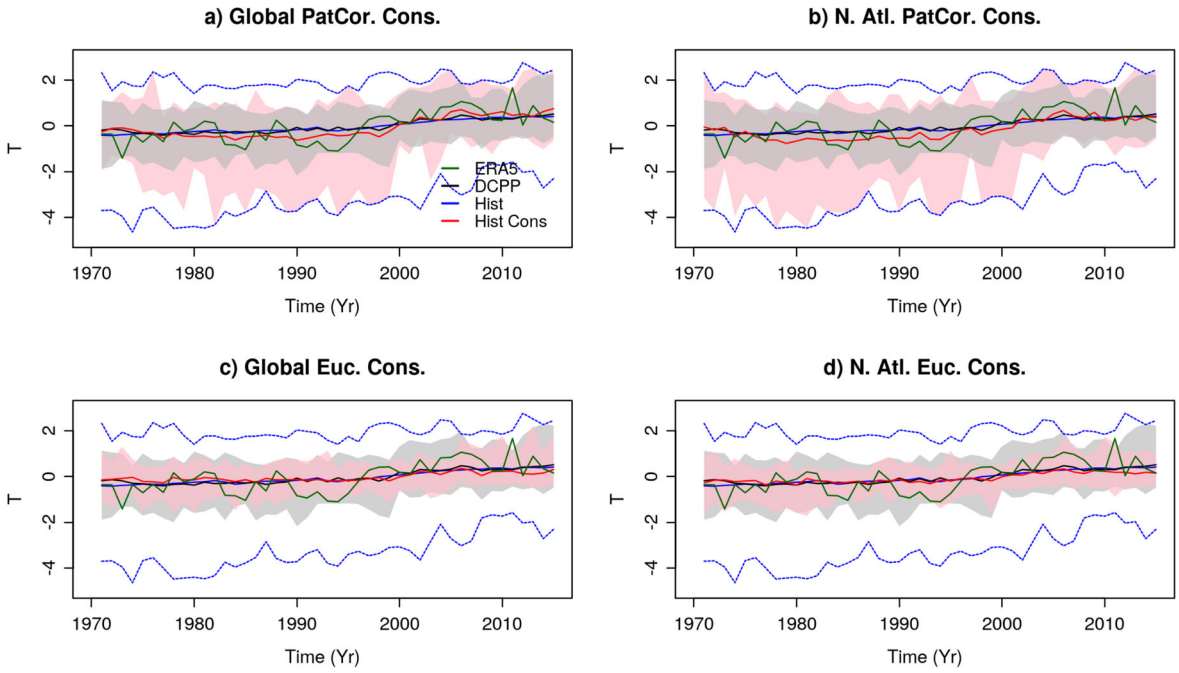


Figure 11. Subpolar North Atlantic SST (°C) at the transition between the decadal predictions and historical simulations for the full and constrained ensembles using four methods. Gray and pink shading are the minimum-maximum range of the decadal prediction and the historical constrained ensembles, respectively. Dashed blue lines are the minimum-maximum range of the full historical ensemble.

The next steps will involve testing the centered root mean square difference as an alternative to the Euclidean distance. This approach could help address issues encountered during member filtering and potentially improve the distribution of the constrained members. We will also test if results improve by expanding the size of the historical ensemble by including all available historical simulations. The larger pool should help identify better analogs, at the expense of introducing simulations from models that are not represented in the decadal prediction ensemble, which could potentially degrade the physical consistency of the blended product. While this analysis has focused on sea surface temperature (SST), we will also apply the blending methodology to near-surface temperature and sea level pressure over Europe, and assess its effectiveness for predicting extreme events for the demonstrator cities that are relevant —taking advantage of the ensemble distribution.

3.3 Progress Beyond State of the Art

This deliverable presented four new innovative blending methods developed in Task 5.2 to provide relevant time-to-event forecasts (NRC) and seamless and relevant climate information over the next decades (DMI, CNRS-Cerfacs, BSC).

3.4 Discussion and Next Steps

Several blending methods for time-to-event forecasts and decadal timescales were developed in Task 5.2 and presented in this deliverable:

- Several blending methodologies were explored to perform time-to-event forecasts (i.e. forecasts that estimate the timing of a specific event) from different numerical weather predictions, using as an example application the first hard freeze of the winter season for locations in Norway and Fennoscandia. These methods generally improve upon single-source forecasts when the sources are balanced in terms of noise-to-signal ratio and there is enough hindcast data. These methods can be used to study other time-to-event forecasts, such as a sudden stratospheric warming or a drought period.

- A new blending method was developed to provide skillful estimates of winter climate over Eurasia for the next ten years, using the observed teleconnection between the NAO and surface air temperature to constrain historical simulations. The evaluation shows that this method could offer skillful winter surface air temperature predictions on multi-annual timescales at minimal cost compared to state-of-the-art initialized decadal climate predictions.

- A new blending methodology explored the potential benefits of combining the three available sources of information —observations, initialized decadal predictions, and non-initialized climate projections— to provide relevant information on near-term climate change with reduced uncertainty related to internal climate variability. Its assessment for the prediction of winter surface temperature over the Mediterranean region as a case study highlights the added value of this method in comparison to the historical ensemble or hindcast prediction alone.

- A new blending method was developed to combine decadal predictions and climate projections to provide seamless climate information for the next 30 years. This method is based on constraining climate projections based on the ensemble of decadal predictions, which allow us to blend predictions and projections, maintaining the statistical properties of both ensembles, making it suitable for probabilistic predictions, including of climate extremes.

These blending methods are promising and their assessments highlight the potential benefits for improving our information about future weather and climate, on timescales that range from a few weeks to several years to decades ahead.

The next step will be to compare these blending methods, in terms of consistency and usefulness for several case studies in Task 5.3.

4 Impact

We will discuss these new methodologies with other colleagues and collaborators in the sister project ASPECT, so we can compare the advantages and disadvantages with respect to other methodologies developed within ASPECT, which could lead to further improvements.

The results from these new methodologies, especially for near-term future prevision, can be of particular interest for WP6 of the I4C project. Further discussions are needed to ensure a good alignment with WP6 needs.

In this way the outcomes of this deliverable will contribute to I4C Expected Outcomes 3 ("Improved assessment of risks for people and systems exposed to extreme weather a climate events") and 4 ("Enhanced scientific collaboration and exploitation of synergies across the EU and Associated Countries").

5 Links Built

There is an ongoing discussion with partners of WP6 to select some target variables of relevance for the demonstrator cities. This will be then included in the case studies to be presented in Deliverable 5.3 assessing the ability of I4C's blending methodologies.

6 Communication, Dissemination and Exploitation

Participation to conferences

- The new methodology developed by CNRS-Cerfacs has been presented in the EGU 2024 (oral presentation), the 15th International Meeting on Statistical Climatology at Toulouse (oral presentation) and the EMS 2024 annual meeting in Barcelona (oral presentation).

Peer reviewed articles:

There are articles by all groups in preparation related to the corresponding work shown in this deliverable.

- Cunen, C., Roksvåg, T., Heinrich-Mertsching, C., & Lenkoski, A. (2025). Combining predictive distributions for time-to-event outcomes in meteorology. **In review**. *arXiv preprint arXiv:2503.19534*.
- Mahmood, R., Yang S., Donat M., (2025), Constraining the NAO-temperature teleconnection in CMIP6 simulations enables skillful multi-annual predictions of Eurasian winter climate, Environmental Research Letters, **in review**.

7 References

Baran, S. & Lerch, S. (2018). Combining predictive distributions for the statistical post-processing of ensemble forecasts. International Journal of Forecasting 34, 477–496. https://doi.org/10.1016/j.ijforecast.2018.01.005

Baker, L. H., Shaffrey, L. C., Sutton, R. T., Weisheimer, A., & Scaife, A. A. (2018). An intercomparison of skill and overconfidence/underconfidence of the wintertime

North Atlantic Oscillation in multimodel seasonal forecasts. *Geophysical Research Letters*, *45*(15), 7808-7817. https://doi.org/10.1029/2018GL078838

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.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., ... & Eade, R. (2016). The decadal climate prediction project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751-3777. https://doi.org/10.5194/gmd-9-3751-2016

Cunen, C., Roksvåg, T., Heinrich-Mertsching, C., & Lenkoski, A. (2025). Combining predictive distributions for time-to-event outcomes in meteorology, https://arxiv.org/abs/2503.19534

Donat, M. G., Mahmood, R., Cos, J., Ortega, P., Doblas-Reyes: Improving the forecast quality of near-term climate projections by constraining internal variability based on decadal predictions and observations. Environ. Res.: Climate in press https://doi.org/10.1088/2752-5295/ad5463 , 2024.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937-1958. https://doi.org/10.5194/gmd-9-1937-2016

Hawkins, E., and Sutton, R., 2009 The Potential to Narrow Uncertainty in Regional Climate Predictions. Bull. Am. Meteorol. Soc., 90 1095–107 Online: http://journals.ametsoc.org/doi/abs/10.1175/2009BAMS2607.1

Hermanson, L., Eade, R., Robinson, N. H., Dunstone, N. J., Andrews, M. B., Knight, J. R., ... & Smith, D. M. (2014). Forecast cooling of the Atlantic subpolar gyre and associated impacts. *Geophysical research letters*, *41*(14), 5167-5174. https://doi.org/10.1002/2014GL060420

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. https://doi.org/10.1002/qj.3803

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., ... & Zhang, H. M. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *Journal of Climate*, *30*(20), 8179-8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (2003). An overview of the North Atlantic oscillation. *Geophysical Monograph-American Geophysical Union*, *134*, 1-36. https://doi.org/10.1029/134GM01

Kalbfleisch, J. D. & Prentice, R. L (2002). The statistical analysis of failure time data. Wiley. https://doi.org/10.1002/9781118032985

Kerr, R. A. (2000). A North Atlantic climate pacemaker for the centuries. *Science*, *288*(5473), 1984-1985. https://doi.org/10.1126/science.288.5473.1984

Kushnir, Y., Scaife, A. A., Arritt, R., Balsamo, G., Boer, G., Doblas-Reyes, F., ... & Wu, B. (2019). Towards operational predictions of the near-term climate. *Nature Climate Change*, *9*(2), 94-101. https://doi.org/10.1038/s41558-018-0359-7

Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., ... & Hawkins, E. (2020). Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth System Dynamics*, *11*(2), 491-508. https://doi.org/10.5194/esd-11-491-2020

Lussana, C. (2020). seNorge observational gridded datasets. seNorge_2018, version 20.05. *arXiv preprint arXiv:2008.02021*. https://doi.org/10.48550/arXiv.2008.021
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, 2022. https://doi.org/10.5194/esd-13-1437-2022

Meehl, G.A., L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F.J. Doblas-Reyes, E. Hawkins, A. Karspeck, M. Kimoto, A. Kumar, D. Matei, J. Mignot, R. Msadek, H. Pohlmann, M. Rienecker, T. Rosati, E. Schneider, D. Smith, R. Sutton, H. Teng, G.J. van Oldenborgh, G. Vecchi and S. Yeager (2014). Decadal climate prediction: An update from the trenches. Bulletin of the American Meteorological Society, 95, 243-267, https://doi.org/10.1175/BAMS-D-12-00241.1

Gneiting, T. & Ranjan, R. (2013). Combining predictive distributions. Electronic Journal of Statistics, 7, 1747–1782. https://doi.org/10.1214/13-EJS823

Sanchez-Gomez, E., Cassou, C., Ruprich-Robert, Y., Fernandez, E., & Terray, L. (2016). Drift dynamics in a coupled model initialized for decadal forecasts. *Climate Dynamics*, *46*(5), 1819-1840. https://doi.org/10.1007/s00382-015-2678-y

Sutton, R. T., & Dong, B. (2012). Atlantic Ocean influence on a shift in European climate in the 1990s. *Nature Geoscience*, *5*(11), 788-792. https://doi.org/10.1038/ngeo1595

Smith, D. M., Eade, R., Scaife, A. A., Caron, L. P., Danabasoglu, G., DelSole, T. M., ... & Yang, X. (2019). Robust skill of decadal climate predictions. *Npj Climate and Atmospheric Science*, *2*(1), 13. https://doi.org/10.1038/s41612-019-0071-y

Stephenson, D. B., Pavan, V., Collins, M. M. J. M., Junge, M. M., Quadrelli, R., & Participating CMIP2 Modelling Groups. (2006). North Atlantic Oscillation response to transient greenhouse gas forcing and the impact on European winter climate: a CMIP2 multi-model assessment. *Climate Dynamics*, *27*, 401-420. https://doi.org/10.1007/s00382-006-0140-x

Trenberth, K. E., & Shea, D. J. (2006). Atlantic hurricanes and natural variability in 2005. *Geophysical research letters*, *33*(12). https://doi.org/10.1029/2006GL026894

Trigo, R., Osborn, T., & Corte-Real, J. (2002). The North Atlantic Oscillation influence on Europe: Climate impacts and associated physical mechanisms. Climate Research, 20, 9–17. https://doi.org/10.3354/cr020009

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., ... & Lovenduski, N. S. (2018). Predicting near-term changes in the earth system: a large ensemble of initialized decadal prediction simulations using the community earth system model. *Bulletin of the American Meteorological Society*, *99*(9), 1867-1886. https://doi.org/10.1175/BAMS-D-17-0098.1

Ye, K., Messori, G., Chen, D., & Woollings, T. (2022). An NAO-dominated mode of atmospheric circulation drives large decadal changes in wintertime surface climate and snow mass over Eurasia. Environmental Research Letters, 17(4), 044025. https://doi.org/10.1088/1748-9326/ac592f