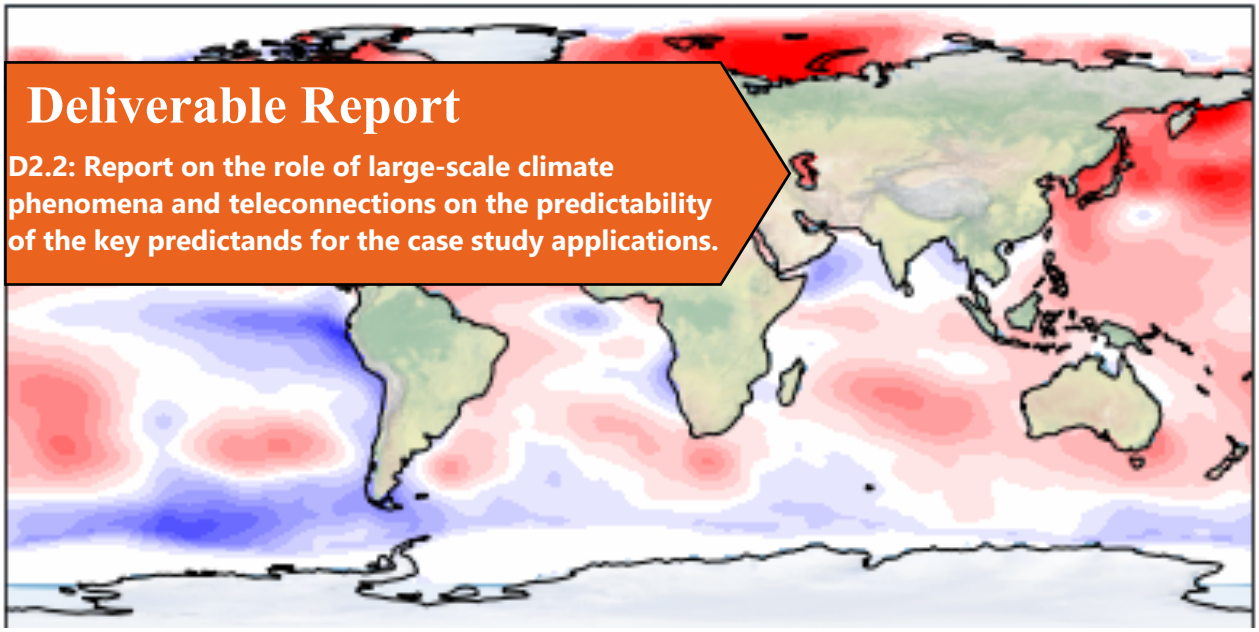

The Added Value of Seasonal Climate Forecasts for Integrated Risk Management Decisions (SECLI-FIRM)

EU H2020 Project (ref. n. 776868)

D2.2: Report on the role of large-scale climate phenomena and teleconnections on the predictability of the key predictands for the case study applications (v2)

Improving the skill of seasonal forecasts through multi-model combination, advanced statistical methods, and signal boosting.

[Dissemination level: Public]

Version Table

Name/Party	Description	Date
K Nielsen (UL)	Draft one	3/1/2021
Victor Estella Perez	Merging all contributions	23/3/2021
Victor Estella Perez	Corrected J.Vidal's comments	29/3/2021
Victor Estella Perez	Corrected A. Troccoli's comments	11/04/2021
Victor Estella Perez	Revisions based on reviewers' feedback	3/08/2021
A Troccoli	Final version, revised and ready for submission	31/08/2021

Internal Review Table

Name/Party	Description	Date
José Vidal (UL)	WP Leader's review	26/3/2021
Alberto Troccoli (UEA)	Project Leader's review	7/4/2021
Alberto Troccoli (UEA)	Project Leader's final review	12/4/2021
José Vidal (UL)	WP Leader's review	4/8/2021

Contributors (Consortium Party/Person)

UEA	Alberto Troccoli, Folmer Krikken, Elena Maksimovich
ENEL	Gaia Piccioni, Antonio M. Nicolosi, Marco Formenton, Elena Calcagni, Elvira Musicò, Gloria Rea, Martina Morgani
ENEA	Franco Catalano
UL	Jose Vidal, Kristian Nielsen, Victor Estella-Perez
KNMI	Gertie Geertsema
WEMC	Luke Sanger
Météo France	Christian Viel
Met Office	Hazel Thornton

Table of Contents

1 Introduction

Seasonal climate forecast has shown the potential for providing useful information in decision making in large parts of the world (Weisheimer and Palmer 2014). However, over Europe, the skill of most seasonal forecast models in both summer and winter are more limited for most variables (Mishra et al. 2018). Moreover, the signal to noise ratio is often very low for the forecast (Scaife and Smith 2018). This also leads to small variance in the signal that can prove difficult to exploit in decision-making to the stakeholder. To develop and test different post-processing tools for improving these aspects, and to help tailoring the forecasts for the stakeholder in the SECLI-FIRM project, several individual studies were carried out by the participating partners in task 2.2. The focus of these studies has been on three different, but related, topics:

a. Multi model and multi model ensemble (MME) combinations for improved skill of both deterministic and probabilistic seasonal and monthly climate forecasts to address the needs of the stakeholders represented in some of the case study (CS) work carried out in WP3 (Section 2.1). In relation to this, a study of a developed metric to estimate the independency between probabilistic forecast from different systems has been performed (Section 2.2). This is intended to provide a prior knowledge about the optimal combination of models along with an explanation about the improvement that this combination presents over other combinations of less independent models.

b. Tree-based regression system has been tested as statistical forecasting systems for seasonal and monthly forecasting of climatic variables to test their skill against the dynamical models in relation to CS5 (Section 2.3). In relation to this study the addition of a single dynamical models was added to develop a hybrid system building on the Random Forest algorithm. Further, a global assessment of the added value of these statistical systems relative to the dynamical forecasting models are tested for potential adoption in several of the CSs (Section 2.4). This was done in a multi model setup including statistical forecasting systems based on observations only and dynamical forecasting models available through the work carried out by Task 2.1.

c. Two different approaches have been tested to boost seasonal forecast signals. The first method is building on a selection of the best performing ensemble members from an MME with respect to the North Atlantic Oscillation (NAO) in Section 2.5. The average of these is used to inflate the signal and achieve better predictions in relation to the root mean square error (RMSE) and Pearson correlation of specific variables against observations than using the prediction from the mean of all ensemble members from the MME.

   The second method focuses on boosting the mean of the predictions also by exploiting probabilistic aspects of the forecast and selecting a subset of ensemble members in cases when the forecast is confident (Section 2.6). The general assumption for this method is that if a forecast shows a likelihood greater than a predefined threshold of a specific event happening, such as warmer or colder than the climatic mean, then the

average of the ensemble is found only using the ensemble member that predicted this. This leads to an inflation of the signal that only depends on how certain the model is that an event will happen at a given time, and not on the actual skill of the model. This results in a boosting of the signal in both the case when the model predicts correct but as well when the prediction is wrong. However, an advantage with this method is that it can be performed without prior knowledge of the outcome of the forecast. This boosting approach was used in Case Studies 1-5 by ENEL, for which the main results are presented in Section 2.7.

Additionally, in Section 2.8, a brief justification of the work delivered by Météo-France on the impact of North Atlantic Weather Regimes in the downscaling process. Despite relying on large-scale climate phenomena and teleconnections, due to its major strength of this approach being the use of weather regimes, Météo-France work presented a better fit to Task 2.3 and is therefore described in detail in deliverable D2.3.

In general, the methods presented showed different amount of success in improving the tailored seasonal forecasts usability for the stakeholders in regard to improving the skill or boosting the signal. Although not all the results presented here have been directly adopted by the SECLI-FIRM case studies during the project execution, the industry users have been involved in presentations and discussions of this work and in some cases they plan to test these procedures after the project. Of those adopted, results from the multi-model combination and independence metric (sections 2.1 and 2.2), as well as the signal boosting method using a simple threshold (section 2.6 and 2.7), have been used in the implementation of ENEL's case studies 1 to 5.

As part of Task 2.2 other lines of research were pursued but due to their non-conclusive results, also due to the resignation of a team member, these results are not documented in this report. This is the case for instance on the research on the teleconnection between the Monsoon-Desert Mechanism and its related prediction signal over Euro-Mediterranean. Initial investigations using the ECMWF SEAS5 were summarised with the first interim report (pp 30-31). While this research had a strong potential in terms of providing a predictive signal for the Mediterranean region with about a month horizon, priority was given to other more focused uses of seasonal forecasts, in the context of the SECLI-FIRM case studies, as discussed in this report.

2 Developed methodologies

2.1 Estimation of the added skill of the optimal multi model combination

All dynamical weather and climate forecast systems are bound to have uncertainties due to their parametrization of physical processes as well as inaccuracies in their methods of solving some physical processes (Lee et al. 2013). This combined with uncertainties in observed and analysed initial conditions and the chaotic nature of weather and to some degree, seasonal climate variations tend to lead to inaccuracy.

To account for this, combining the ensembles from the independent skilful seasonal forecast systems into a multi model ensemble (MME) has proven to be a reliable method to improve the skill of these forecasts (Palmer et al. 2005). The reason is that the widening of the ensemble spread achieved by including different models helps mitigate the overconfidence of individual forecast systems. Furthermore, error cancellation between models may also help in minimising the error of the mean of the MME compared with the individual models.

This however raises the question: Do all models always provide the best forecast with the highest skill or are there specific model combinations that provide a more significant improvement? In this work, we aimed to test all individual combinations and thereby estimate the potential improvement in skill between using the best combination compared to that of all models together as well as individually. This was done for both specific forecasts related to case studies in WP3, as well as for larger, more general areas. The variables chosen for investigation was 2-metre temperature and total precipitation.

2.1.1 Methodology

A similar approach as the one by Alessandri et al. 2018 was chosen to evaluate the potential of an optimal combination of models to improve the skill compared to utilizing all models available. This was done by calculating the skill of all combinations to find the most skilful combination. All the forecasts are seasonal, and the focus has been on DJF and JJA forecasted from November and May with a lead time of one month.

Two main skill scores have been chosen for the analysis of the combination, the anomaly correlation coefficient (ACC), and the Brier skill score (BSS). These were chosen to account for both deterministic and probabilistic approaches used in different case studies and therefore different skill scores were needed to test the combinations. The ACC is widely used to assess the quality of seasonal forecasts when utilizing the mean of a model ensemble. Here we are using it to evaluate the quality of the ensemble mean from the models for each combination. In its simplest form, the formula for calculating the ACC is written out in equation 2.1.1.

$$ACC = \frac{\overline{(f-c)(o-c)}}{\sqrt{\overline{(f-c)^2}\ \overline{(o-c)^2}}} \qquad \text{(Eq. 2.1.1)}$$

Where *f* is the forecasted value of the variable, *c* is the climatic mean of the month being forecasted and *o* is the observed value of the variable.

The Brier Skill Score allows estimating the improvement of the accuracy of a probabilistic forecast against a reference forecast. It is defined in equation 2.1.2 as

$$BSS = 1 - \frac{BS_{for.}}{BS_{ref.}}. \qquad \text{(Eq. 2.1.2)}$$

Where the Brier Score for the forecast ($BS_{for.}$) and for the reference forecast system which the forecast is tested against ($BS_{ref.}$) are calculated following equation 2.1.3

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \qquad \text{(Eq. 2.1.3)}$$

Where N is the number of ensemble members, f_t is the predicted probability at time *t* from the forecast that an event will happen and o_t is the observation of whether the event happened. Following this, o_t takes values of 0 or 1 (0 if the event was not observed and 1 if the event was observed). This is similar to the mean squared error for deterministic evaluations, however, following Murphys (1973) decomposition of the metric, it was shown that it includes both reliability, resolution and uncertainty of the forecast. Therefore, it is widely used as a general evaluation of the skill of probabilistic seasonal climate forecasts.

For this probabilistic forecast, the events chosen for investigation are the lower and upper terciles for both precipitation and temperature. These variables are of interest to many of the end-users and play and significant role for both energy demand and for some specific cases such as case study 3 (CS3) and CS5 for energy production potentials. It would have been interesting to test even more extreme events; however, the limited number of available years for analysis limits how extreme an event can be investigated. To quantify if a cold/dry event (observation under the threshold for the lower tercile) or warm/wet event (observations larger than the threshold for the upper tercile) occurred, the seasonal terciles are determined from the climatic distribution of the average of the 3 months under investigation using the observational information from ERA5. For each model, the tercile thresholds are determined compared to the climatic distribution of the models forecast with a lead time of one month. This ensures that even with biases in the mean or variance of the forecast, the predicted probability for these specific events to occur from the models are comparable with the observations from ERA5.

The likelihood of an event happening is evaluated as the number of ensemble members predicting it divided by the total number of ensemble members chosen. To ensure an equal and fair estimation of skill, 10 ensemble members of each model are used, as this is the number of members available from the model with the smallest ensemble.

The combination of the forecast systems has been done giving equal weight to each system in this study, this allowed to create a general method that could be tested by different case studies and used for a case-by-case tailored choice of models for an MME.

The areas chosen for investigation, and illustrated in Figure 2.1.1, are East-Europe (35-70 ^{o}N, 15-40 ^{o}E), West-Europe (38-55 ^{o}N, 10 ^{o}W-18 ^{o}E), Mediterranean (35-47 ^{o}N, 10 ^{o}W-35 ^{o}E), Italy (35-47 ^{o}N, 7-18 ^{o}E) Spain (35-44 ^{o}N, 10 ^{o}W-3 ^{o}E) and Colombia (6 ^{o}S-15 ^{o}N, 83-65 ^{o}W). These regions provide larger geographical areas for more statistically significant testing and still cover many of the case studies.

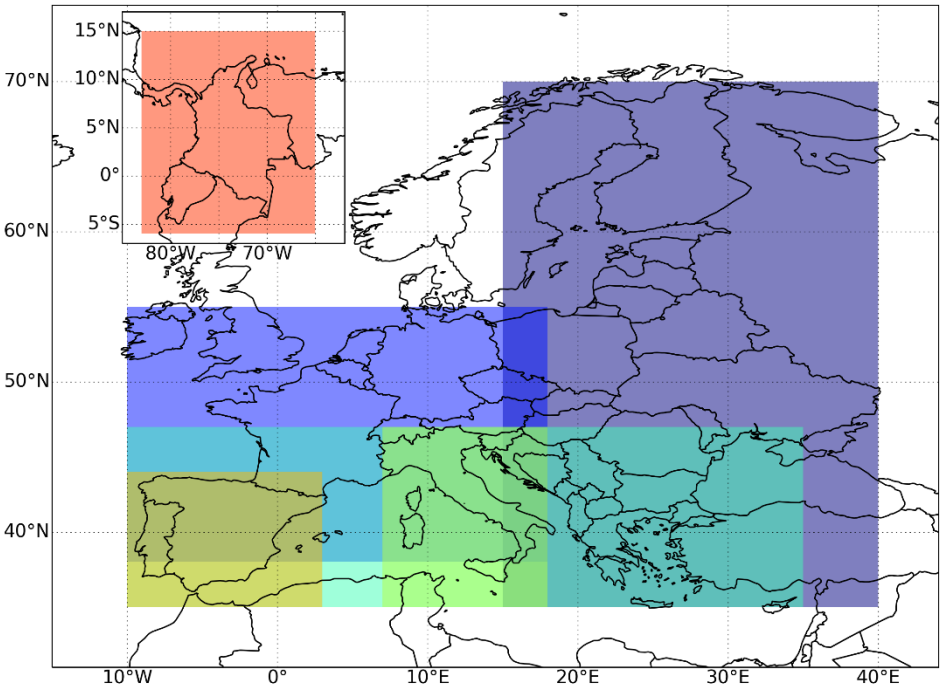

Figure 2.1.1: Geographical illustration of the domains chosen. Colombia (orange shaded), Spain (shaded yellow), West-Europe (shaded blue), Italy (shaded green), Mediterranean (shaded teal) and East-Europe (shaded purple)

The score for each season is calculated by finding the score for each combination at each point and then averaging this over all the grid points within the area of interest. In this study, we are focusing on land points only. This is defined by the land see mask from ERA5, only using grid points with values larger than zero.

2.1.2 Data used

From all the hindcast datasets retrieved in the project, a specific selection of independent models was chosen. This meant that different versions of the same models are left out and as a main rule only the newest version is used. This allows for faster computation times and ensures that the versions of the models tested are also available for future incorporation into

the end user's decision processes. The hindcast period covers the start data of all months from 1993-2016 covering a period of 24 years. All models are retrieved at a 1°x1° regular lat/lon grid. For calculating the probabilistic score of a combination 10 ensemble members were selected from each model to ensure a fair weighting between the models. The final selection of 11 different models is shown in table 2.1.1.

Table 2.1.1: Overview of the 11 different models chosen.

Model acronym	**System name**	**Atmos. model**	**Ocean model**	**Members (hindcast/ forecast)**	**Initialization Atmos.**	**Initialization Ocean hind/forecast**
CANI	CanCM4i	CanAM4	CanOM4	10/10	CMC	ORAP5 ocean reanalysis
CCSM	COLA-RSMAS-CCSM4	CAM4	CCSM POP2	10/10	CFSR	OISST
CMCC	CMCC-SPS3.5	CESM 1.2 - CAM 5.3	NEMO v3.4	40/50	ERA5	C-GLORS Global Ocean 3D-VAR
DWD	GCFS 2.1	ECHAM 6.3.05	MPIOM 1.6.3	30/50	ERA5	ORAS5
ECMWF	SEAS5	IFS Cycle 43r1	NEMO v3.4	25/51	ERA-Interim	ORAS5
GEMN	GEM-NEMO	GEM 4.8-LTS.13,	NEMO 3.6 ORCA	10/10	ERA-interim	ORAP5 ocean reanalysis/
GFDL	SPEAR	AM 4.0	MOMv6	15/30	CFSR	OISST v2
JMA	JMA/MRI-CPS2	JMA-GSM	MRI.COM v3	10/13	JRA-55	MOVE/MRI.COM-G2
MF	System 6	ARPEGE v6.2	NEMO v3.4	25/71	ERA-interim	GLORYS2V2 / Mercator-Ocean
NCEP	CFSv2	GFS	GFDL MOM4	24/120	CFSR	CFSR
UKMO	GloSea5-GC2-LI	Unified Model (UM) - Global Atmosphere 6.0	NEMO v3.4 - Global Ocean 5.0	28/60	ERA-Interim	GS-OSIA / FOAM

ERA5 was used as the observational information to assess the skill of the MME as well as the single model forecast. This was retrieved from Copernicus Data Store (CDS) at a 1x1 regular lat/lon grid to fit with the resolution of the models. The benchmark forecast for the BSS was chosen as climatology, meaning there is a 1/3 chance of exceeding the tercile for the given event.

2.1.3 Results - Deterministic

The focus of some of the case studies have been on a tailored deterministic forecast for precipitation, therefore the MME combination method is tested using the ACC for the areas relating to CS3 and CS5 (specific catchment areas of ENEL and Celsia in Italy and Colombia respectively, see D3.3 and D3.6). Figure 2.1.2 illustrates the testing of combinations for a deterministic forecast of temperature over areas of interest to ENEL in Italy. The forecast is monthly with a lead of one month. The correlation for each combination was therefore found between 288 months of forecast and observations (Feb 1993 to Jan 2017).

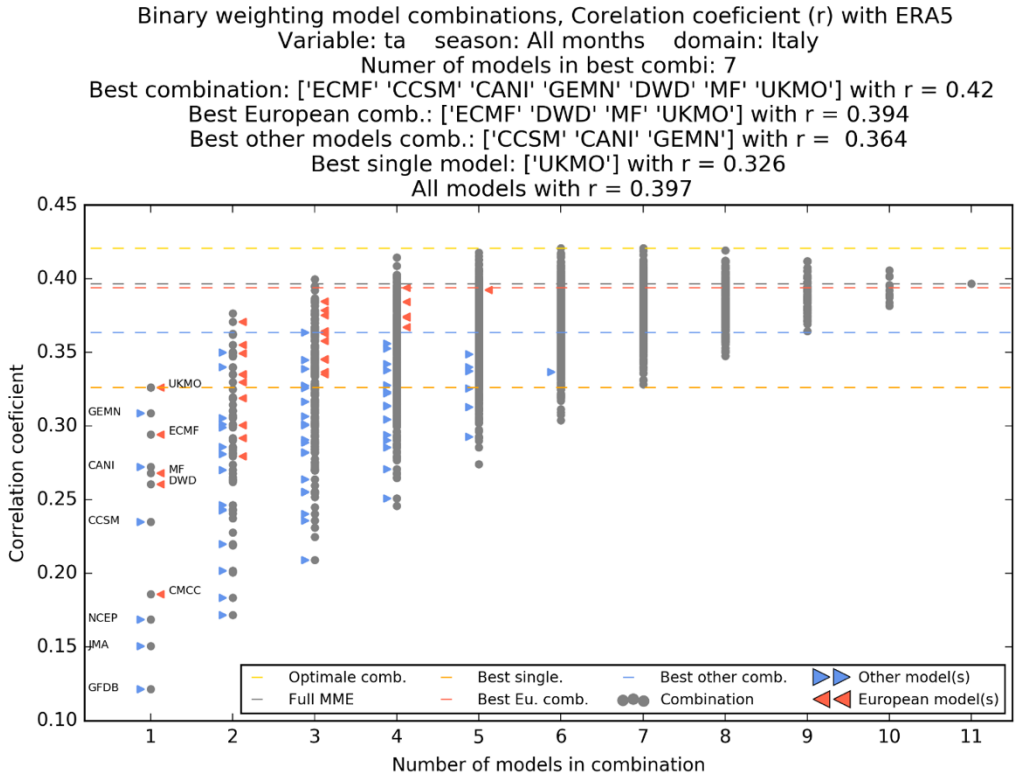

Figure 2.1.2: Results of the MME combinations method for a deterministic forecast of temperature over areas of interest to ENEL in Italy. Grey dots represent the score from each combination. Red triangles represent the cases in which the combinations are obtained with only models from the European community, while blue triangles are for the combinations of the non-European models only. The combinations mixing models from both the European, the NMME and the JMA communities are the grey dots without marking.

For this case, the best combination had an R-value of 0.42 and the r-value for all models were 0.40 with the specific improvement of 0.023. To test if this improvement is significant a bootstrap method was used as a statistical test of significance. All 110 ensemble members from the 11 models are collected in a pot. Then as the best combination consisted of 7 models, 70 ensemble members are randomly picked, while allowing replacement from the pot containing all members. The correlation between the average of these and the observations is

found. Similar 110 members are selected randomly, while allowing replacement, and the correlation between the average of these and the observations is found. Finally, the difference between these two correlation coefficients is found. This procedure is repeated 10.000 times to form a distribution. From this distribution, the confidence intervals were found. The left plot in Figure 2.1.3 shows the distribution of the correlation between the mean of the randomly sampled ensembles and observations. Furthermore, the distribution of the differences between the correlation coefficients is shown in the right plot of Figure 2.1.3. This leads to the conclusion the improvement in ACC of the best combination over using all models in this case only is statistically significant to the 80% significant level (Table 2.1.2).

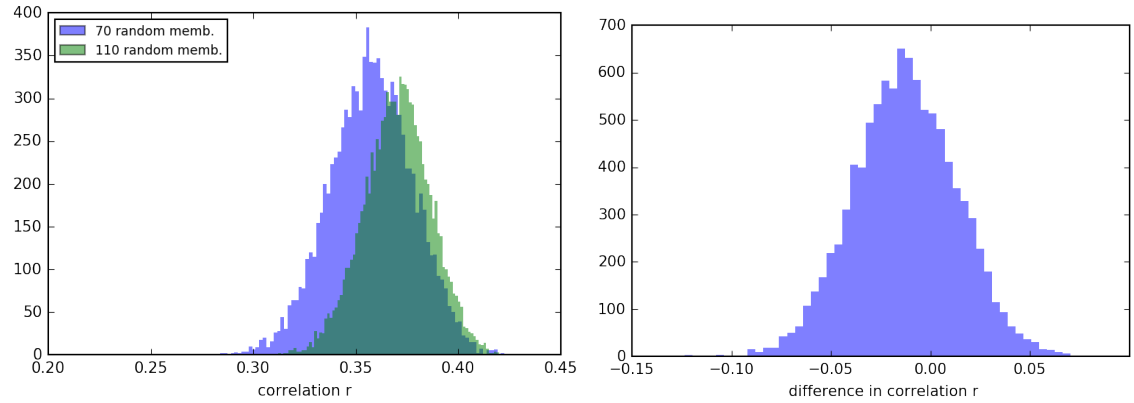

Figure 2.1.3: Output of the statistical testing of significance for the forecast of temperature relating to CS3. Left plot: Distribution of the ACC for the 70 (blue bins) and 110 (green bins) randomly sampled ensemble members with repetition from the full pot of all possible members. Right plot: Distribution of difference between the ACC for the

Table 2.1.2: Confidence intervals for the statistical testing of significance of the forecast of temperature relating to CS3.

Confidence interval	60%	70%	80%	90%	95%	99%
minimum	-0.034	-0.040	-0.046	-0.055	-0.063	-0.081
maximum	0.0009	0.015	0.021	0.030	0.038	0.054

For CS5 a test of a monthly forecast with one month lead time of temperature over the 4 grid points covering Bogota (75-74 ºW, 4-5 ºN) was carried out as it is assumed that temperature is a driving factor of energy consumption in the city. It is shown that there is in general a higher skill of the forecast as temperature anomalies in this area are strongly linked to the El Niño /La Niña phenomena as illustrated in Figure 2.1.4.

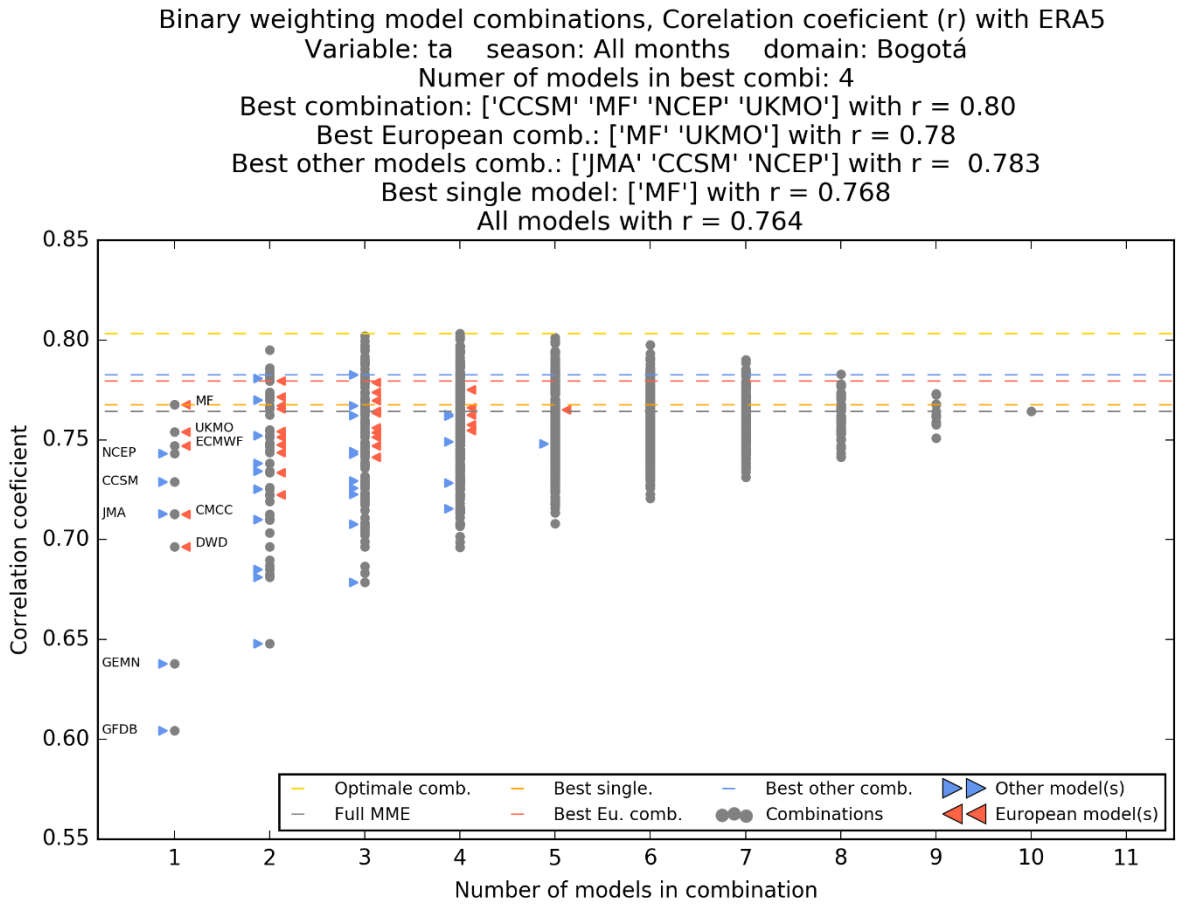

Figure 2.1.4: Same as Figure 2.1.3 but for the domain covering Bogota.

The significance test of this area showed that there is a significant improvement of 0.039 by using the best model combination with a correlation coefficient value of 0.80 compared to that of all models with a correlation coefficient of 0.76 with a 95% certainty as the uncertainty range for this confidence level covers –0.034 to 0.027 see Table 2.1.3.

Table 2.1.3: Same as Table 2.1.2 but for the domain covering Bogota.

Confidence interval	60%	70%	80%	90%	95%	99%
minimum	-0.016	-0.019	-0.023	-0.029	-0.034	-0.044
maximum	0.010	0.013	0.017	0.022	0.027	0.036

This indicates that there is a benefit in using the best combination of models for this specific area and variable in relation to ACC for the forecast of the average of the ensemble members compared to using all the tested models as a combined full MME.

2.1.4 General results and conclusions for probabilistic forecast over European domains.

The results of the larger domains can be seen in Table 2.1.4 where for each seasonal forecast covering DJF and JJA with a lead time of 1 month the BSS for the best single model, overall best combination and the combination of all models is reported. The predicted variables were TP and TA for the upper and lower terciles.

Table 2.1.4: Summary of the results of the testing of the MME method for temperature over the selected domains with respect to the BBS of the seasonal forecast of winter (DJF) and summer (JJA) season. The scores are noted for the Best individual model, the combination of all models and the best combination overall. The number of models in the best combinations is shown in parentheses and significant improvements are indicated with bold.

JJA 2m Temperature lower tercile				**DJF 2m Temperature lower tercile**			
	Best individual	*All models*	*Best combination (nr. mod)*		*Best individual*	*All models*	*Best combination (nr. mod)*
Colombia	0.42	0.36	0.46 **(3)**	Colombia	0.39	0.34	0.4 **(4)**
EastEU	0.15	0.19	0.21 **(5)**	EastEU	0.16	0.15	0.2 **(4)**
Italy	0.17	0.2	0.24 **(3)**	Italy	0.08	0.09	0.13 **(3)**
MediEU	0.17	0.22	0.23 (5)	MediEU	0.11	0.9	0.13 **(4)**
Spain	0.15	0.14	0.18 **(4)**	Spain	0.14	0.11	0.16 **(2)**
WestEU	0.08	0.11	0.13 **(7)**	WestEU	0.14	0.1	0.16 **(6)**
JJA 2m Temperature upper tercile				**DJF 2m Temperature upper tercile**			
	Best individual	*All models*	*Best combination (nr. mod)*		*Best individua*	*All models*	*Best combination (nr. mod)*
Colombia	0.28	0.29	0.37 **(3)**	Colombia	0.46	0.5	0.51 (6)
EastEU	0.1	0.16	0.18 **(6)**	EastEU	0.17	0.17	0.2 **(3)**
Italy	0.17	0.17	0.25 **(3)**	Italy	0.15	0.16	0.21 **(4)**
MediEU	0.16	0.2	0.23 (6)	MediEU	0.13	0.15	0.17 (5)
Spain	0.15	0.14	0.17 **(3)**	Spain	0.17	0.16	0.21 **(4)**
WestEU	0.14	0.09	0.14* **(1)**	WestEU	0.14	0.15	0.17 (5)

Table 2.1.5: Same as table 2.1.4 for precipitation.

JJA Precipitation lower tercile				**DJF Precipitation lower tercile**			
	Best individual	*All models*	*Best combination (nr. mod)*		*Best individual*	*All models*	*Best combination (nr. mod)*
Colombia	0.15	0.2	0.24 **(3)**	Colombia	0.23	0.26	0.29 **(3)**
EastEU	0.06	0.11	0.12 (7)	EastEU	0.09	0.12	0.14 (3)
Italy	0.06	0.09	0.11 (4)	Italy	0.09	0.13	0.14 (5)
MediEU	0.05	0.12	0.13 (6)	MediEU	0.10	0.13	0.15 **(6)**
Spain	0.11	0.16	0.18 **(4)**	Spain	0.17	0.17	0.22 **(3)**
WestEU	0.05	0.09	0.11 (5)	WestEU	0.1	0.13	0.15 (5)
JJA Precipitation upper tercile				**DJF Precipitation upper tercile**			
	Best individual	*All models*	*Best combination (nr. mod)*		*Best individual*	*All models*	*Best combination (nr. mod)*
Colombia	0.14	0.21	0.23 **(4)**	*Colombia*	0.22	0.25	0.28 **(5)**
EastEU	0.06	0.12	0.12 (7)	*EastEU*	0.08	0.12	0.13 (5)
Italy	0.08	0.1	0.13 **(3)**	*Italy*	0.1	0.12	0.14 **(5)**
MediEU	0.08	0.13	0.14 (7)	*MediEU*	0.09	0.13	0.14 (7)
Spain	0.1	0.16	0.17 (5)	*Spain*	0.16	0.16	0.18 **(4)**
WestEU	0.04	0.11	0.12 (7)	*WestEU*	0.08	0.11	0.14 **(4)**

From these results (Tables 2.1.4 and 2.1.5) it can be seen that often (≈63% of the time) there is a combination of models that is significantly more skilful than the combination utilizing all models. Furthermore, this seems to occur more often for more skilful forecasts of temperature (≈79% of the time) than the less skilful forecasts of precipitation (≈46% of the time).

Having this knowledge allows users to not only have a slight increase in forecast skill, but also a smaller number of models needed to achieve the optimal combination, decreasing the work related to gathering the data and working with them. This can be seen as an additional benefit of doing this prior selection of the optimum forecast models.

However, a case-by-case estimation is still needed to ensure that there is a benefit from using the ensemble from the best combination instead of the full ensemble from all available models. This testing for case-specific areas also reveal that in some less frequent cases for these tested domains a decrease in skill by using all available models with equal weight over the single most skill full model is observed. An example of this can be seen for a forecast of DJF temperatures in the lower tercile in the Colombian domain, see Figure 2.1.5.

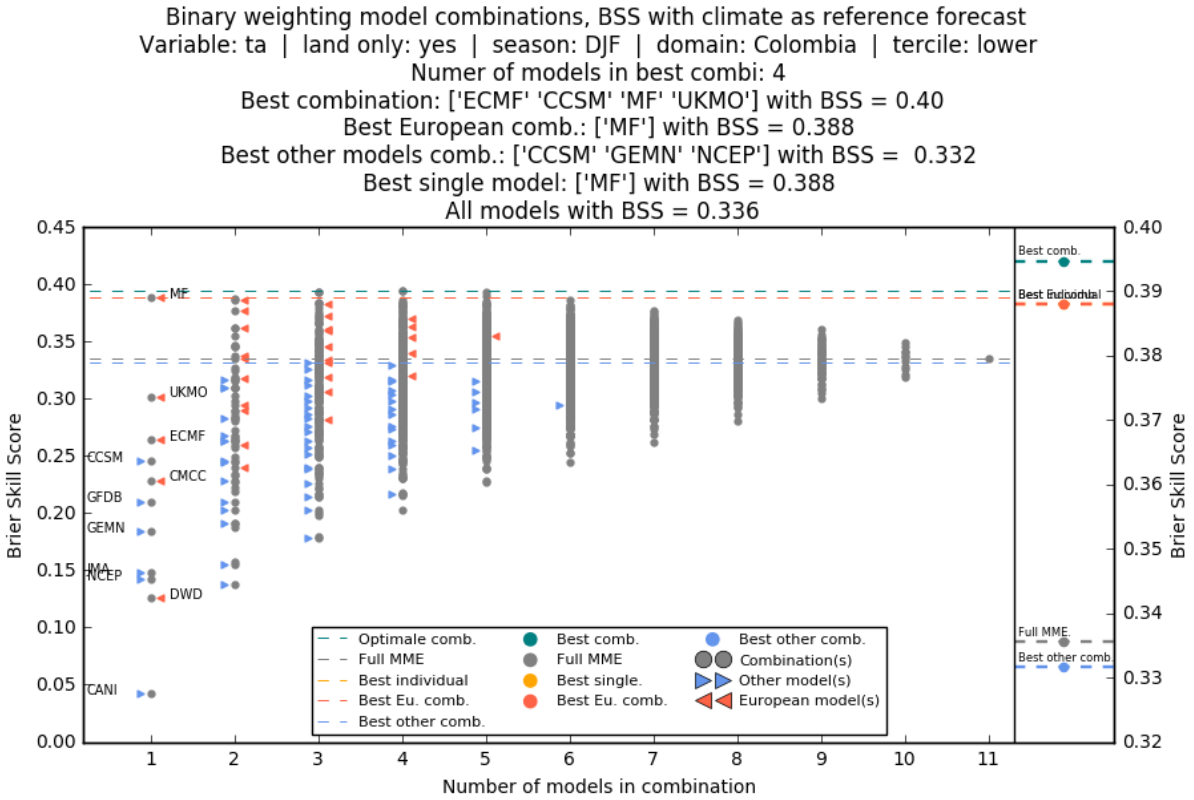

Figure 2.1.5: Results of the MME method tested for a seasonal forecast for DJF with a lead time of one month for temperature in the lower tercile over the Colombian domain. Grey dots represent the score from each combination. Red triangles represent the cases in which the combinations are obtained with only models from the European community, while blue triangles are for the combinations of the non-European models only. The combinations mixing models from both the European, the NMME and the JMA communities are the grey dots without marking.

But for the forecast of TP in JJA in Spain the opposite is basically true as almost all combinations of models perform better concerning the BSS than any single model as illustrated in Figure 2.1.6.

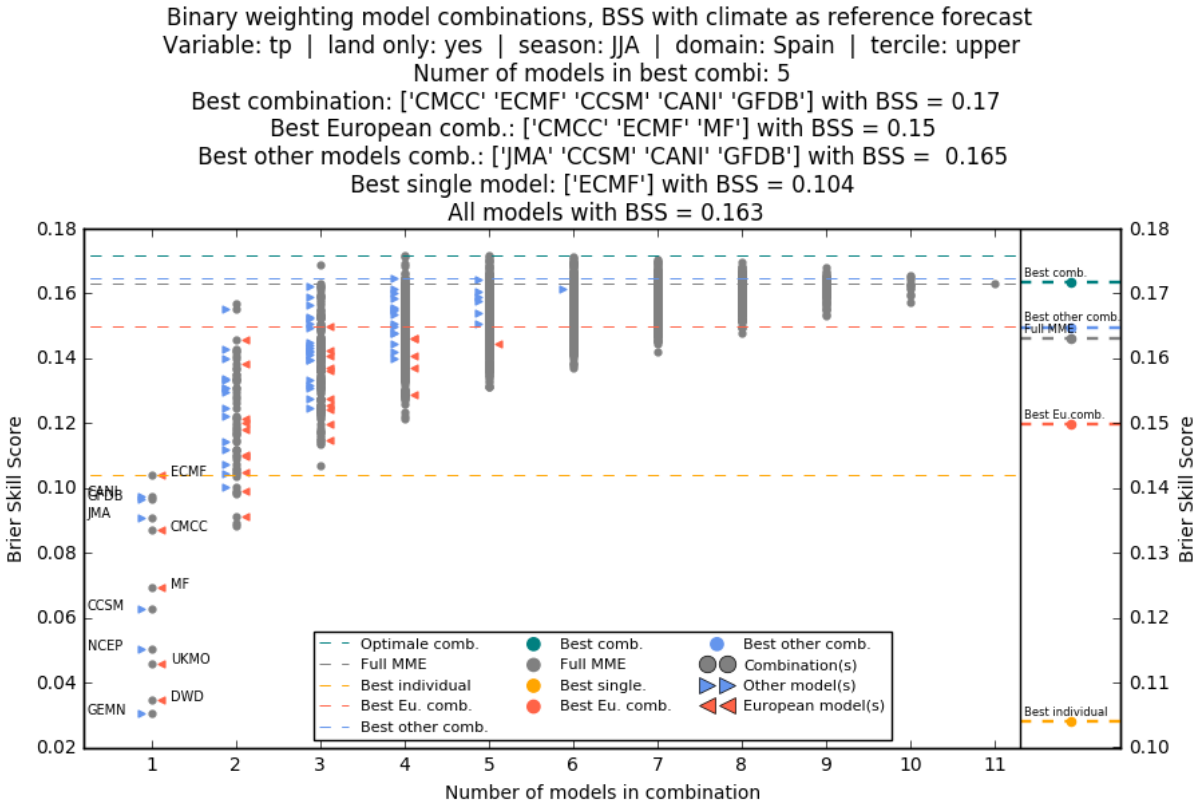

Figure 2.1.6: Results of the MME method tested for a seasonal forecast for JJA with a lead time of one month for precipitation rates in the upper tercile over the Spain domain. Grey dots represent the score from each combination. Red triangles represent the cases in which the combinations are obtained with only models from the European community, while blue triangles are for the combinations of the non-European models only. The combinations mixing models from both the European, the NMME and the JMA communities are the grey dots without marking.

These examples further underline the significant differences in the process of improving the skill from MME combinations regarding the geographical area of the domain, the variable, the season, and the general skill of the individual forecast models. It is therefore a priori hard to suggest any selection of models for an optimal combination. It has been proven in an earlier study by Allesandri et al. 2017 that combining more independent models with different physical representations of atmospheric and ocean processes results in a higher skill. Therefore, being able to evaluate the independence between different models could provide a priori knowledge of what models would perform best when combined with each other as opposed to taking all the available models. Such a metric has been developed and tested utilizing results from this analysis in SECLI-FIRM and is described under section 2.2.

2.2 Importance of probabilistic independence for MME combination optimization

Multi-model ensembles (MMEs) are powerful tools in dynamical climate prediction as they account for the overconfidence and the uncertainties related to single model ensembles. The potential benefit that can be expected by using an MME amplifies with the increase of the independence of the contributing seasonal prediction systems (Alessandri et al., 2018). To this aim, we have collected and analysed prediction systems from the Copernicus C3S seasonal forecasts product (https://climate.copernicus.eu/seasonal-forecasts), the North American Multi-Model Ensemble and the Japan Meteorological Agency (Table 2.2.1) (see D2.1).

One-month lead retrospective seasonal predictions are collected for the considered models for the period 1993-2017 (1st May and 1st November start dates, i.e., June-July-August, JJA and December-January-February, DJF). The validation period is limited to 2014 for the analysis involving surface albedo due to the availability of satellite observations of GLCF-GLASS data (Liu et al., 2013). On the other hand, ERA5 reanalysis (Hersbach et al., 2018) is the reference dataset for all the other surface climate variables considered. We analysed the seasonal hindcasts in terms of deterministic scores (anomaly correlations and its decomposition in yearly normalized covariance) and probabilistic score (Brier Skill score) with a particular focus on land domain, since little evaluation has been performed so far over land domains that is where a large number of applications of seasonal forecasts are based. New metrics are developed in order to assess the relative independence of the prediction systems in the probabilistic information they provide. The multi-models get their performance from the skill of the contributing models, so that MME skill is generally proportional to the mean skill of the individual models. However, the relation between single-model averages and MME skill is not linear and the multi-model performance is superior to the average of the single-model ensembles mainly because of error cancellations. The independence of the contributing models between each other is a prerequisite to obtain error cancellations and for skill amplification to occur (Hagedorn et al., 2005).

2.2.1 Process-based model inter-comparison

In order to show how different models contribute to MME performance, the deterministic skill (anomaly correlations, ACC) of a subset of models from C3S (ECMWF, Météo-France and DWD) has been compared together with associated possible predictability sources.

Overall, the ECMWF prediction of T2M in DJF is outperforming the MF prediction over East Europe and Central Asia while the Météo-France predictions tend to be better over East US and West Europe (Figure 2.2.1a). There is a pronounced negative skill difference for winter (DJF) surface temperature in parts of the North Atlantic in the ECMWF model with respect to Météo-France. Over that region there is a known problem in the ECMWF system related to the ocean initialization with Ocean Reanalysis System 5 (ORAS5, Johnson et al., 2019). The affected region is centred on a box defined by the longitudes 50-30 °W and the latitudes 45-55 °N and it can potentially affect forecasts over Europe through advection by the prevailing westerly winds. Indeed, the comparison of ACC shows that the 1-month lead forecasts initialized 1st November tend to have less skill in predicting 2m temperature over West Europe.

Surface temperature prediction in the winter season is strongly related to the representation of snow-albedo processes while surface solar radiation variability is affected by both local surface conditions (evapotranspiration) and the atmospheric dynamics through moisture convergence (Alessandri et al., 2017). To investigate the coupling and the possible predictability sources, the relationships between the improvement of the correlation for the target variables (e.g. 2 m-temperature and surface solar radiation) is analysed with respect to the improvements in the possible drivers for the areas of interest (e.g. surface albedo, moisture convergence). For this purpose the correlation coefficient is decomposed in its components measuring the covariance between each predicted (x) and observed (y) yearly (i) anomalies [$r(x,y)_i$ hereinafter normalized yearly covariance,], following the approach in Alessandri et al. (2017). The model 1 minus model 2 difference in the normalized yearly covariance [$\Delta r(x,y)_i$] is analyzed to identify the possible driver contributor to the enhanced predictability of the target variables resulting from the different model and/or initialization strategies. To this aim, the linear relation between $\Delta r(x,y)_i$ of the target and driver fields is assessed using a least square method and significance of the slope of linear relationship is evaluated using a Fisher parametric test. The positive linear relationship between target and driver in terms of the model 1-minus-model 2 $\Delta r(x,y)_i$ indicates the change of predictability of the target as mediated by the driver, which is directly affected by the differences in the two prediction systems. Only the linear coefficients of the regression that passed significance test at 10% level are considered. The analysis revealed a strong local coupling of the increased skill in 2m temperature, over East Europe coming from the snow processes represented by surface albedo (Figure 2.2.1b). Positive (negative) values of normalized yearly covariance differences mean better (worse) skill in system 1 with respect to system 2 in predicting the driver and target variables. Indeed, the fact that most of the years occur in the upper right quadrant indicates that increases in the prediction of surface albedo also drives enhancement of T2M forecasts.

The same analysis has been applied to compare ECMWF and DWD systems (Figure 2.2.2a). There are large differences between the two models, in particular over continental areas. Here the DWD model performs better over the Iberian Peninsula, West Europe, and most of Asia (except for India, Indonesia, and Japan), while the ECMWF model shows better correlations over Canada, South America, and Africa. The ECMWF model, in turn, gives better predictions over Canada, Indian monsoon region and Sahel. The two models share the same ocean initialization strategy (ORAS5) and therefore do not show significant differences over the North Atlantic. The normalized yearly covariance differences scatterplot (Figure 2.2.2b) shows, again for the East EU domain, that the skill difference of 2m temperature is consistently related to the ability of the model to represent land surface albedo processes.

The comparison of ECMWF vs Météo-France for 1-month lead seasonal hindcasts for boreal summer (June-July-August, JJA) surface solar radiation (Figure 2.2.3a) shows that the ECMWF model is performing better over North America, East Europe, and Central Asia while Météo-France is giving larger skill over Central Europe, North Africa, China, and East Asia. Interestingly, there are still some negative differences over the North Atlantic similarly to DJF. The analysis revealed the influence of the atmospheric dynamics on the skill via a significant relation between surface solar radiation normalized yearly covariances to moisture convergence over Central Europe domain (Figure 2.2.3b).

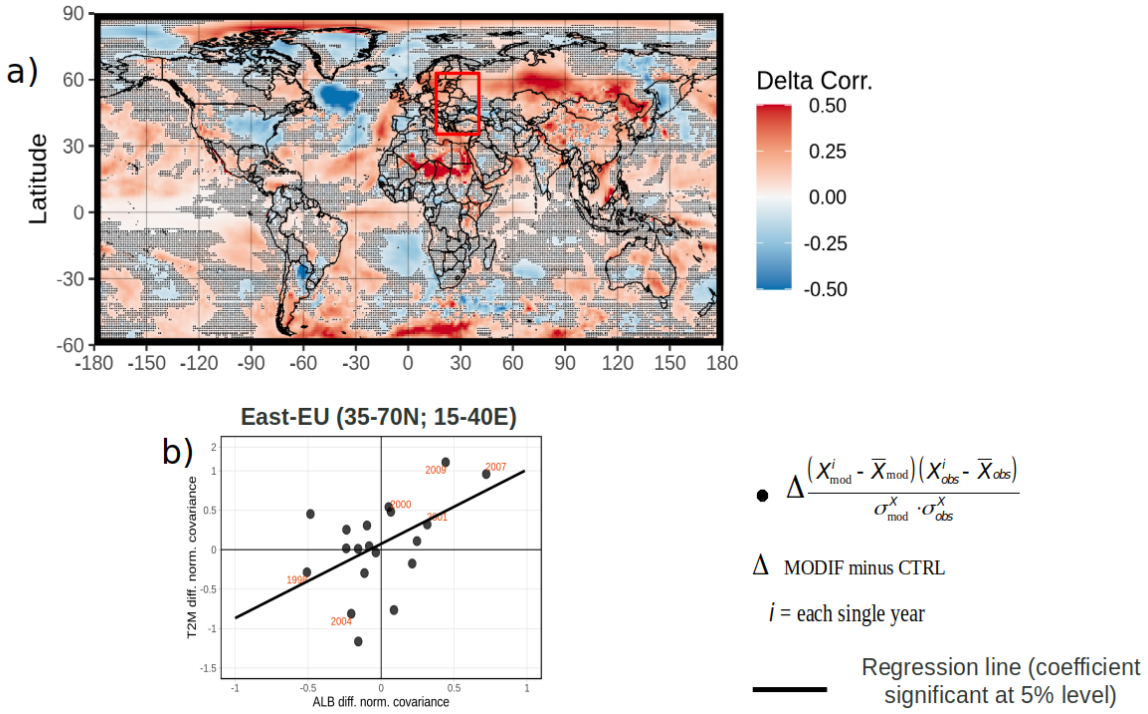

Figure 2.2.1. a) 1-month-lead boreal winter (DJF) 2m temperature ECMWF minus Météo-France correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and Météo-France for the predictions averaged over the East-European domain (15E–40E; 35N–70N) of T2M versus albedo. Black filled circles are the normalized yearly covariance differences computed for each start date. Regression line indicates significant (10 % level) relationship between prediction of target T2M and driver albedo. Orange years indicate when normalized yearly covariance difference change in the same direction (i.e. both target and driver lying in the lower/upper terciles of their respective distribution).

The differences between ECMWF and DWD systems for surface solar radiation in JJA are shown in Figure 2.2.4a. ECMWF gives better predictions over Central Europe, Central Asia, the Amazon, Sahel, and Central Africa while DWD is better over North America, East Russia, and North Africa. Over Central Europe there is still a strong relationship between surface solar radiation and moisture convergence (Figure 2.2.4b) but in this case the influence of the moisture convergence on the surface solar radiation appears to be better represented in the ECMWF than in DWD system.

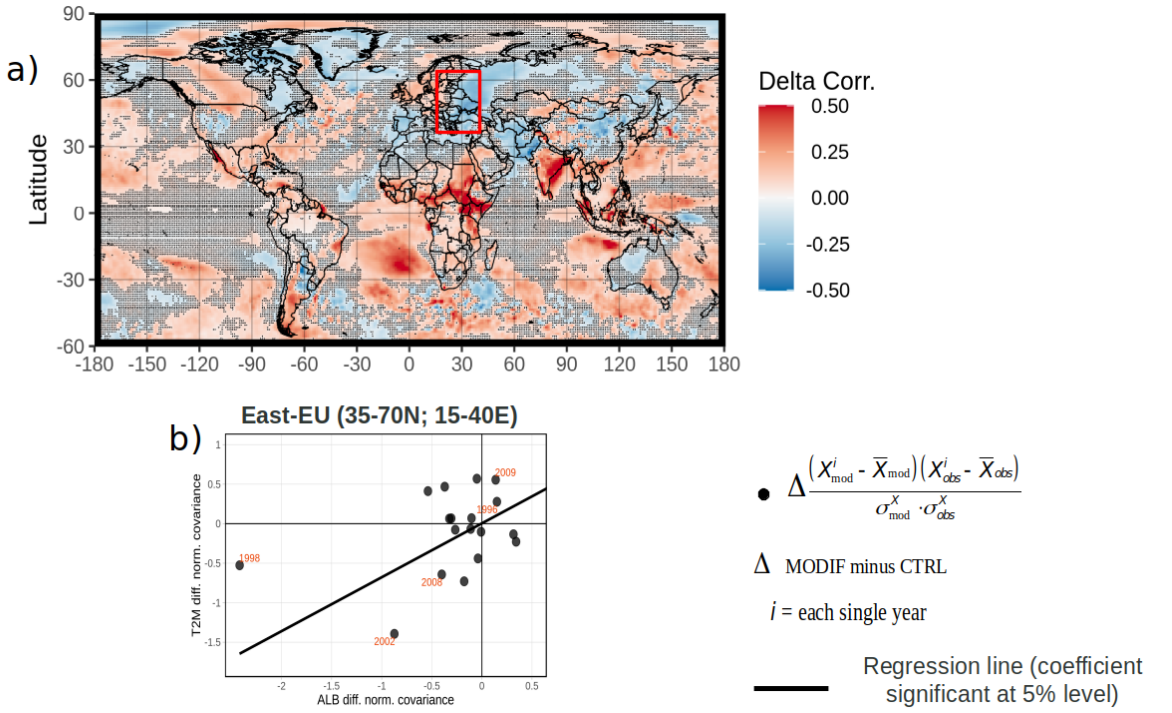

Figure 2.2.2. a) 1-month-lead boreal winter (DJF) 2m temperature ECMWF minus DWD correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and DWD for the predictions averaged over the East-European domain (15 ºE–40 ºE; 35 ºN–70 ºN) of T2M versus albedo. Black filled circles are the normalized yearly covariance differences computed for each start date. Regression line indicates significant (10 % level) relationship between prediction of target T2M and driver albedo. Orange years indicate when normalized yearly covariance difference change in the same direction (i.e. both target and driver lying in the lower/upper terciles of their respective distribution).

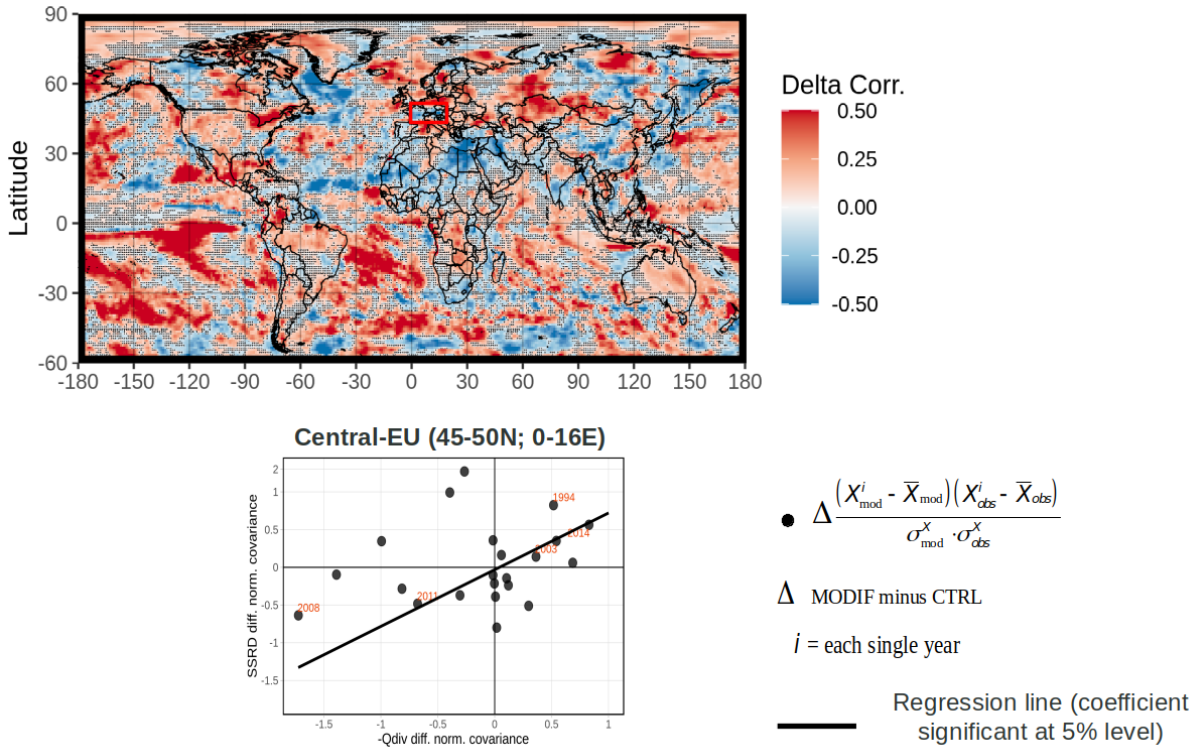

Figure 2.2.3. a) 1-month-lead boreal summer (JJA) surface solar radiation downward ECMWF minus Météo-France correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and Météo-France for the predictions averaged over the Central-European domain (0 ºE–16 ºE; 45 ºN–50 ºN) of SSRD versus moisture convergence. Black filled circles are the normalized yearly covariance differences computed for each start date. Regression line indicates significant (10 % level) relationship between prediction of target SSRD and driver -Qdiv. Orange years indicate when normalized yearly covariance difference change in the same direction (i.e. both target and driver lying in the lower/upper terciles of their respective distribution).

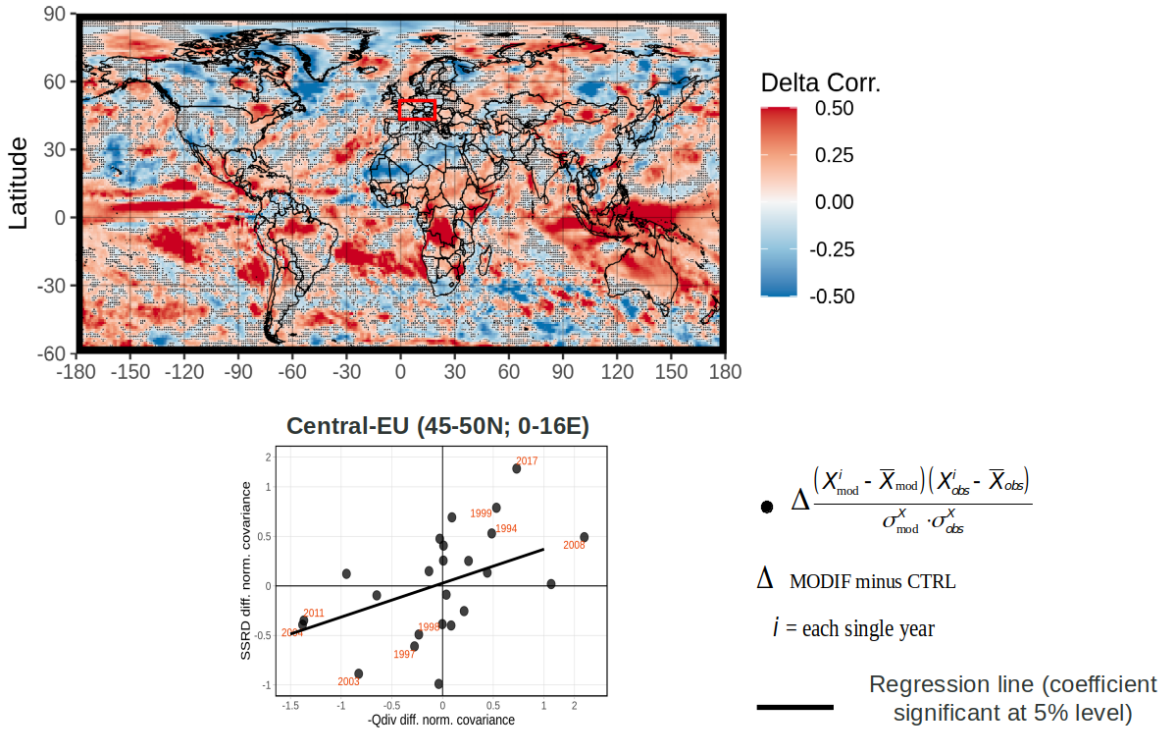

Figure 2.2.4. a) 1-month-lead boreal summer (JJA) surface solar radiation downward ECMWF minus DWD correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and DWD for the predictions averaged over the Central-European domain (0 °E–16 °E; 45 °N–50 °N) of SSRD versus moisture convergence. Black filled circles are the normalized yearly covariance differences computed for each start date. Regression line indicates significant (10 % level) relationship between prediction of target SSRD and driver -Qdiv. Orange years indicate when normalized yearly covariance difference change in the same direction (i.e. both target and driver lying in the lower/upper terciles of their respective distribution).

2.2.2 *Probabilistic scores and model independence*

The probabilistic accuracy has been analysed in terms of Brier Skill score (BSS) for dichotomous events of conditions being above (below) upper (lower) tercile of the sample distribution. Furthermore, starting from the definition of the Brier score (Equation 2.1.3; Wilks, 2011), we have developed the Brier score covariance (BScov; Equation 2.2.1) metric (see also S2S4E D4.4), which estimates the relative independence of prediction systems 1 and 2:

$$BS_{cov} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_{1,i} - o_i\right)^2\left(y_{2,i} - o_i\right)}{\sqrt{BS_1 \cdot BS_2}} \qquad \text{(Eq. 2.2.1)}$$

where i indicates each hindcast year and n total number of years; y is forecast probability and o is for the observed [0, 1] dichotomous event under consideration. Subscripts (1) and (2) in Equation 2.2.1 indicate system 1 and 2, respectively. The aim of the new metric is to provide quantitative information on the relative independence of the prediction systems and therefore guidance on the best combination strategies for the selection of the models contributing to the MME. BS_{cov} is equal to 1 when the two systems are the same (system1 = system 2) and its value decreases with increasing model independence. Due to the fact that, by definition, BS_{cov} considers both inter-model distance and distance with respect to observations, the values tend to be concentrated towards its upper limit.

For this part of the analysis, we focused on the European domain. Results for 2m temperature BSS for the lower tercile in DJF are mostly consistent with the analysis of the deterministic scores. The larger positive skill differences between ECMWF and Météo-France are concentrated over East Europe and in general at the higher latitudes while Météo-France system is performing better over the Iberian Peninsula and the Mediterranean countries (Figure 2.2.5a). The comparison of ECMWF with DWD confirms the better performance of the latter system over continental areas and in particular on the Eastern part of the domain (Figure 2.2.5b). The BS_{cov} metric has been used to assess the relative independence of the selected models in the probabilistic information they provide (Figure 2.2.5c and d). For both combinations, the larger probabilistic independence (lower BS_{cov} values) is over the ocean, indicating that model or initialization differences in this component play a major role that must be considered in MME model selection. Over land, the three systems show larger independence over East EU, suggesting that the representation of the snow-albedo processes and the land-surface initialization in the different systems, as discussed in previous section, are important factors to consider for model combination. Interestingly, both for land and the ocean, some regions with small or non-significant skill differences are characterized by large independence (the Mediterranean Sea and Central Europe).

As shown in Figure 2.2.6, models characterized by large relative independence in 2 m temperature (Figure 2.2.6a) also display large independence in surface albedo (Figure 2.2.6b) which has been identified in the analysis above as an important driver for temperature prediction in boreal winter over the East-EU region. This indicates that, in particular for this region, differences in land-surface processes representation (here snow/albedo) are highly contributing to model independence. Furthermore, here we see that adding a model from the NMME community (NCEP) to the European MME ensemble adds a large contribution in terms of model independence.

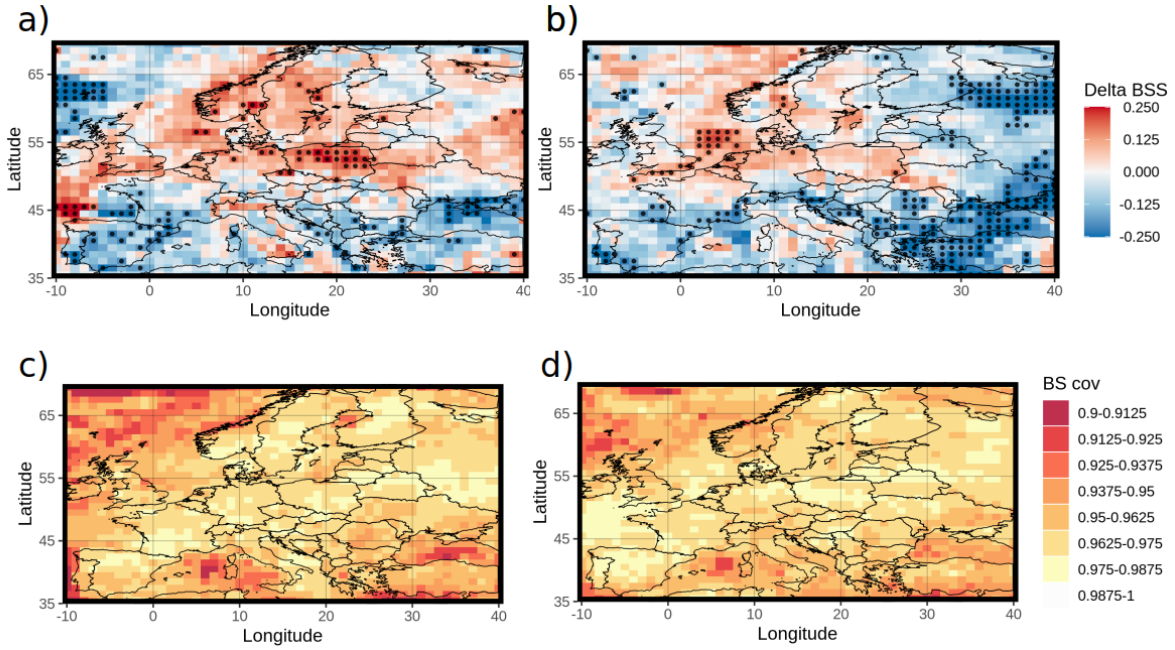

Figure 2.2.5. Spatial distribution of the BSS differences of the probabilistic forecasts for below-normal (below lower tercile of sample distribution) 2m temperature in Boreal winter (DJF) for Europe domain. (a) ECMWF minus Météo-France; (b) ECMWF minus DWD; dotted are the areas that passed a significance test at the 10% level. Probabilistic independence as measured by the new BScov metric: (c) ECMWF vs Météo-France; (d) ECMWF vs DWD.

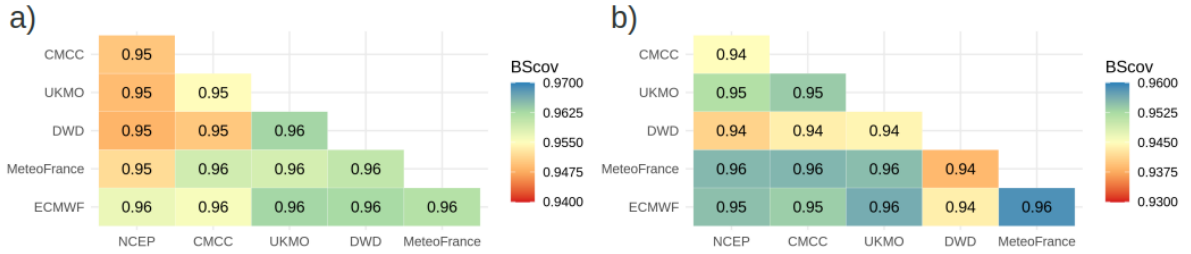

Figure 2.2.6. Probabilistic independence measured by the new BScov metric for East-EU region (35-70 °N; 15-40 °E) for (a) 2m temperature and (b) surface albedo for boreal winter (DJF). Only the models available in Copernicus C3S are considered since albedo is not available for the other NMME models at present time.

Figure 2.2.7 shows the probabilistic scores for the lower tercile of the distribution for surface solar radiation in JJA. In terms of skill, the ECMWF system is performing slightly better than Météo-France over land at the higher latitudes while the latter model is outperforming over Central Europe (Figure 2.2.7a), consistently with the deterministic analysis in previous section. Comparison of ECMWF vs DWD (Figure 2.2.7b) shows positive skill differences over Central

and North France, Western part of Germany, South Italy, and Greece, while the DWD model has higher BSS over Spain, South France, North Italy, and Romania. In terms of model independence (Figure 2.2.7c and d), again we see a large contribution of the ocean component. The fact that ECMWF and Météo-France share the same ocean model but have different ocean initialization strategies suggests that the impact of ocean initialization can be even larger than the differences in the ocean model itself in terms of systems independence. Over land, large signal comes from Central and East Europe for both the model combinations. Again, large degree of independence is also present over regions which are not characterized by significant skill differences. This supports the added value given by this new metric if included in the process of selection and combination of the contributing prediction systems in the MME.

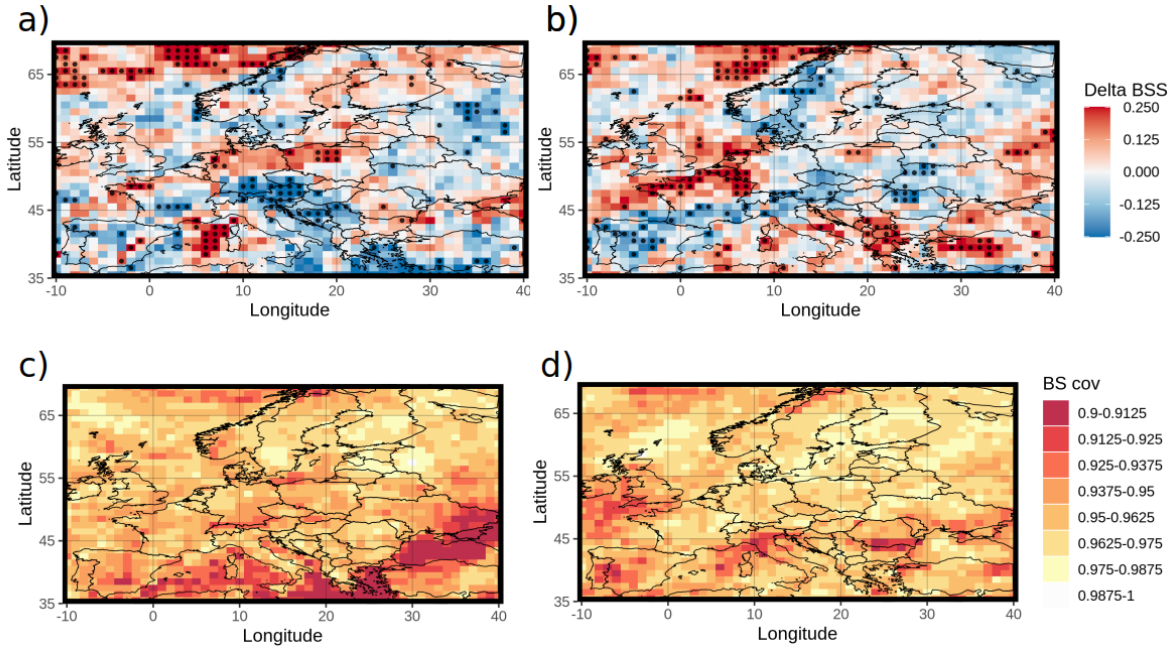

Figure 2.2.7. Spatial distribution of the BSS differences of the probabilistic forecasts for below-normal (below lower tercile of sample distribution) surface solar radiation downward in Boreal summer (JJA) for Europe domain. (a) ECMWF minus Météo-France; (b) ECMWF minus DWD; dotted are the areas that passed a significance test at the 10% level. Probabilistic independence as measured by the new BScov metric: (c) ECMWF vs Météo-France; (d) ECMWF vs DWD.

2.2.3 *Optimization of the MME combination*

In order to assess the maximum level of skill that is currently attainable for seasonal predictions, we have combined the seasonal prediction systems independently developed by the European (ECMWF, CMCC, UKMO, DWD, Météo-France), the North American Multi-Model Ensemble (GEM, CAN, CCSM, GFDL, NCEP) communities, plus the JMA system from the Japan Meteorological Agency into a grandMME consisting of 11 systems. To this aim, all the possible MME combinations have been evaluated by putting together the different systems using equal weights for each model. Figure 2.2.8 shows the BSS averaged over the East-EU

(35-70 °N; 15-40 °E) region for 2 m temperature in DJF as a function of the number of models and obtained with all the possible combinations of the models available. Red triangles represent the cases in which the combinations are obtained with only models from the European community, while blue triangles are for the combinations of the non-European models only. The combinations mixing models from both the European, the NMME and the JMA communities are the grey circles. The maximum performance (yellow dashed line) obtained by mixing models from the European and non-European communities considerably improves what would be obtained by European models only (red dashed line) or by non-European models only (blue dashed line). Figure 2.2.9 shows the relative independence of all the seasonal prediction systems. The best combination identified in Figure 2.2.8 (CMCC, DWD, UKMO, NCEP) corresponds in Figure 2.2.9 to models with high degree of independence from each other.

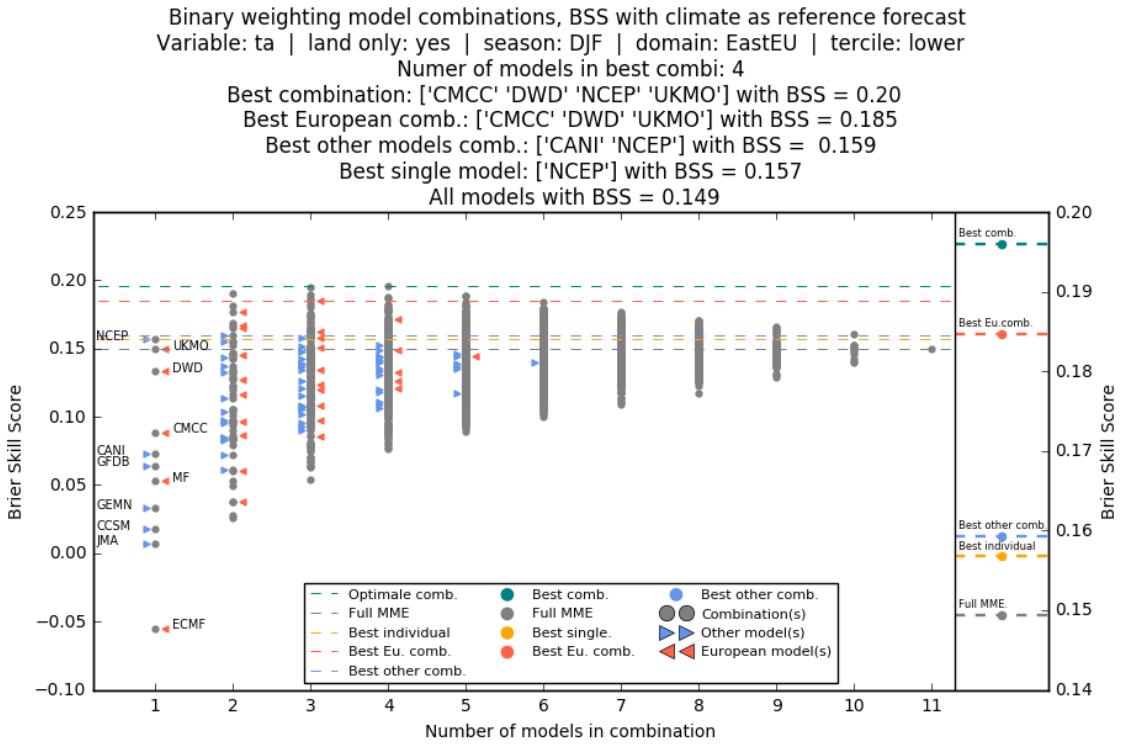

Figure 2.2.8. East-EU (35-70N; 15-40E) Brier Skill Score for boreal winter (DJF) 2 m temperature computed as a function of the number of models obtained with all the possible combinations. Red triangles represent the cases in which the combinations are obtained with only models from the European community, while blue triangles are for the combinations of the non-European models only. The combinations mixing models from both the European, the NMME and the JMA communities are the grey circles

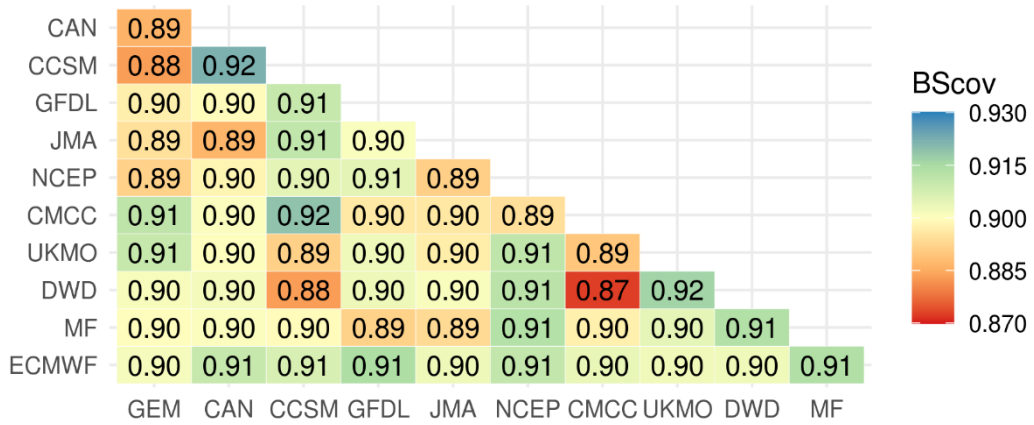

Figure 2.2.9. Probabilistic independence for all the seasonal prediction systems measured by the new BS_{cov} metric for East-EU region (35-70 ºN; 15-40 ºE) for 2m temperature for boreal winter (DJF).

2.2.4 Conclusions

One of the main results of this work, summarised with Figure 2.2.10, is that the skill of the MME combinations increases with increasing degree of independence of the contributing models.

We have developed a methodology to optimize the selection of seasonal forecast models using the MME. The added value of using independence information is the possibility to reduce the number of models and data to produce the optimized forecasts. The methodology developed is general and can be applied for all the variables, seasons, and regions of interest for the energy industry partners.

Two scientific papers are in preparations describing the methodology for the optimization of the MME combination by using model independence information (Alessandri et al. [in preparation], Catalano et al. [in preparation]).

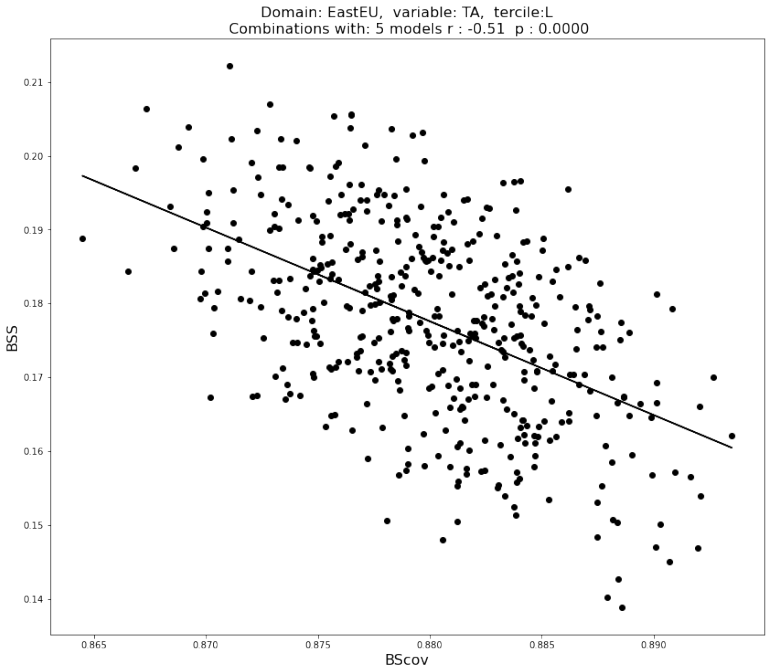

Figure 2.2.10. Scatterplot of the skill vs the average probabilistic independence (larger values of BScov indicate less independence) among the models for all the possible combinations of 5 models for East-EU region (35-70N; 15-40E) for 2m temperature for boreal summer (JJA).

2.3 Random Forest method to enhance the signal of a seasonal forecast system

This work is developed for the industrial user of Case Study 5, Celsia (second component of the CS5). We aim to improve the dam water level predictions currently made by Celsia, which rely on past observations, by adding seasonal forecast information. We explore the use of various methods combining past observations, seasonal predictions and a combination of the two.

The novelty of this work is the use the random forest algorithm (Breiman 2001) to improve the prediction of a variable, the dam water state, based on climatic indices and past observations of the dam. This method creates non-linear relationships between the predictors and predictand (dam state) by fitting various classifying decision trees on randomized subsets of the dataset to define the various parameters of the tree (maximum depth of the tree, minimum number of samples required to be at a leaf node, etc.). Once the characteristics of the tree are defined, we have a division of the phase space of the predictors defined by the different branches of your tree. Predictions are then performed by averaging the predictand of neighbouring points in phase space of the predictors.

We present a description of the data used (2.3.1), the various methods of prediction used with the goal of improving Celsia's predictions (2.3.2) and the results of the predictions (2.3.3). As the predictions that Celsia provided are only between November 2016 and August 2020, we considered two periods for testing our methods: one being 1993-2016 to understand the long-term ability to predictions when various phases of El Niño/La Niña are considered, and a second one, November 2016-August 2020, when we have available Celsia's predictions for comparison but no strong El Niño/La Niña events occur.

2.3.1 Data used

The monthly dam data provided by Celsia ranges from 1947 until 2020 depending on the site. Five sites were provided: Salvajina, Dígua, Anchicaya, Prado and Calima. Anomalies are computed with respect to the climatological period 1993-2016 coinciding with the period of the seasonal forecast data.

All the climate indices used in this study are described in Table 2.3.1, as well as a relationship of their provenance and references. We have considered indices based on the sea surface temperature (SST), atmosphere (ATM) and sea ice extension (SI). The climatology is taken as the average between 1993 and 2016 (both years included) and all of time series of climate indices are computed as anomalies with respect to this climatology.

Predictors are computed using data from various sources. These mainly come from climate indices, previous dam states and predictions from the seasonal prediction system. For past observations we use the last 10 months of the indices shown in Table 2.3.1. We considered 10 months to allow the model to select potential teleconnections that develop within these time scales i.e. Atlantic-Pacific (Polo et al. 2015 or Rodríguez-Fonseca et al. 2009). For observational dam data, we relied on observations of the last 3 months and some statistics of these values (mean and trend). Finally, for the predictions of the seasonal forecast system we

computed with lead times of 1 to 5 months the climate indices indicated with * in Table 2.3.1. Additionally, indices marked with ** in Table 2.3.1 are also computed for the seasonal forecast system but with different definitions from the historical time series. The Atlantic Multidecadal Oscillation (AMO) is taken as the spatial average in the Atlantic between (0,70)ºN and the Pacific Decadal Oscillation (PDO) has been computed as the spatial average (30, 50)ºN, (-180, -140)ºE. The last two indices in Table 2.3.1, local precipitation, and wind speed, are computed only with the Seasonal prediction system data. These indices are computed taking the spatial average of the grid points within a plus/minus 0.5º from the latitude and longitude of the dam location.

Table 2.3.1 – description of the climate indices used as predictors

Index	Component	Origin
El Niño 1+2 *	Ocean	Index from HadISST [-10, 5]ºN - [-90, -80]ºE
El Niño 3 *	Ocean	Index from HadISST [-5, 5]ºN - [-150, -90]ºE
El Niño 3.4 *	Ocean	Index from HadISST [-5, 5]ºN - [-170, -120]ºE
El Niño 4 *	Ocean	Index from HadISST [-5, 5]ºN - [150, -160]ºE
El Niño Modoki *	Ocean	Index from HadISST A – 0.5*(B+C) with A = [-10, 10]ºN - [140, -165]ºE, B = [-15, 5]ºN - [-110, -70]ºE and C = [-10, 20]ºN - [125, 145]ºE
Tropical North Atlantic (TNA) *	Ocean	Index from HadISST [0, 15]ºN - [-80, 0]ºE
Atlantic Niño (AN) *	Ocean	Index from HadISST [-3, 3]ºN - [-40, -20]ºE
Atlantic interhemispheric SST *	Ocean	Index from HadISST A – B with A = [0, 15]ºN - [-80, 0]ºE and B = [-15, 0]ºN - [-80, 0]ºE
Indian Ocean Dipole (IOD) *	Ocean	Index from HadISST A – B with A = [-10, 0]ºN - [90, 110]ºE and B = [-10, 10]ºN – [50, 70]ºE
Atlantic Multidecadal Oscillation (AMO) **	Ocean	NOAA: https://psl.noaa.gov/data/timeseries/AMO
Pacific decadal oscillation (PDO) **	Ocean	NOAA: https://psl.noaa.gov/pdo
Arctic Oscillation (AO)	Atmosphere	NOAA: https://www.ncdc.noaa.gov/teleconnections/ao
Pacific-North American index (PNA)	Atmosphere	NOAA: https://www.ncdc.noaa.gov/teleconnections/pna
North Atlantic Oscillation (NAO) *	Atmosphere	NOAA: https://www.ncdc.noaa.gov/teleconnections/nao
Southern Oscillation Index (SOI) *	Atmosphere	NOAA: https://www.ncdc.noaa.gov/teleconnections/enso/indicators/soi
Arctic Sea Ice extent	Sea ice	NSIDC: https://nsidc.org/data/G02135/versions/3
Antarctic Sea Ice extent	Sea ice	NSIDC: https://nsidc.org/data/G02135/versions/3
Local Precipitation	Atmosphere	Seasonal prediction systems. Computed as the average of (- 0.5,+0.5) for latitude and longitude of the dam location
Local wind speed	Atmosphere	Seasonal prediction systems. Computed as the average of (- 0.5,+0.5) for latitude and longitude of the dam location

The seasonal prediction data have been taken for this study from the European Centre for Medium-Range Weather Forecasts System 5 (Johnson et al. 2019). We have considered only the ensemble mean of the 25 or 51 members (depending on the date) for each of the time series considered.

2.3.2 Methodologies

This work is posed in the way it would be delivered to the end user of Case Study 5. Which means that the goal is to predict the next 1-to-5 months from a particular month using available data up to the previous month (from climate and dam data) and seasonal predictions of next 1-to-5 months. We assumed that the state of current month for both climate and dam is not yet known.

The set-up is depicted in Figure 2.3.1, indicating the span of months for each set of available data that the algorithm uses as predictors: last 10 months for observational climate indices, last 3 months for dam data and predicted indices from the seasonal forecast models at the required lead for the prediction of the dam.

Under these assumptions, we have a set of predictors and a time series to be predicted, the state of the dam with lead 1-to-5 months, which are the input for our random forest model. Due to the variety of time span of the available data, we developed various models combining the possible data. The models are described in Table 2.3.2 with all the predictors (covariates) used in each of the models and time availability of the data. We have used the random forest algorithm with five models combining past climate indices from different sources (SST only - ObsOnly (SST), SST and atmospheric indices - ObsOnly (SST+ATM) and all indices ObsOnly (All)), indices from seasonal prediction system (SeasPred) and a combination of the two data sources (Obs+SeasPred). We considered also a method predicting the dam state with a linear regression algorithm using as predictor the local precipitation from seasonal predictions of each site. Finally, we included also an algorithm used traditionally by energy providers based on past similar situations or analogs (Analogs, later described).

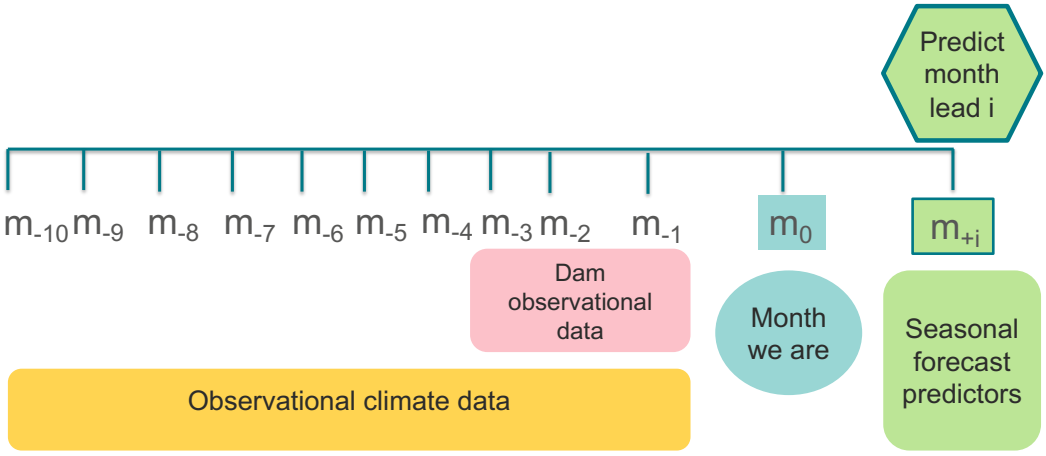

Figure 2.3.1 – Schematic of the time-dependency of the variables used for prediction.

In general, for all the models there is a general structure of the algorithm. This structure follows the steps indicated next:

1. Obtain all required data as predictor and store them in a table
2. Loop for each of the dates that I want to predict:
   a. Create the training data by removing conflicting dates (as the examples explained before)
   b. Train the model with the training data
   c. Obtain a reduction of the predictors (Only for the Random Forest model)
   d. Re-train the model with the selected predictors (Only for the Random Forest model)
   e. Predict with the model the date that we want to predict.

Table 2.3.2 – description of the various models developed

Model Name	Short name	Past observations Predictors	Seasonal System Predictors	Available dates
Analog	Analog	Dam data		As available dam data
Ordinary Least squares	OLS		Precipitation only	From 1993
Only SST indices	ObsOnly (SST)	SST indices + Dam data		From 1900 or as dam data
SST+ATM indices	ObsOnly (SST + ATM)	SST indices + ATM indices + Dam data		From 1960 or as dam data
All indices (SST+ATM+SI)	ObsOnly (All)	SST indices + ATM indices + SI indices + Dam data		From 1980 or as dam data
Seasonal Prediction mean	SeasPred	Dam data	SST indices + ATM indices	From 1993
Past observations and Seasonal Prediction mean	Obs+SeasPred	SST indices + ATM indices + SI indices + Dam data	SST indices + ATM indices	From 1993

Random Forest

The availability of the data used as predictor defines the number of data points from which the random forest can be trained. With less training points the model will not perform well, as not enough branches will be defined by the algorithm. In the case of including the seasonal prediction data as predictors, we can only consider data to train the model from 1993. Limiting our training set with about 300 datapoints, which is not so much regarding than in some cases the number of predictors can reach up to 200 in the case of the model including both past observations and seasonal prediction indices (Obs+SeasPred). In order to take advantage of

most of the available data to be used as predictor we combined on the one hand the use of the Boruta_py algorithm (Kursa et al. 2010) which is described later. On the other hand, we followed two different strategies:

i. *Leave-one-out* – in which we predicted dates between January of 1993 until December of 2016 (both included). We consider as training dataset all the available dates that do not have any information of the current month that we are doing the prediction. For example, if we are predicting January 2000 with lead 3 (with the goal of predicting April of 2000) we would exclude from the dataset the current date and also the data points of February, March, and April of 2000 as the dam state of January is used as predictor for these dates. And also, October 1999 as the dam state is January 2000 is the value we are trying to predict.

ii. "Real-time" Predictions – in this case we predicted between January of 2017 until December 2019 by considering as training data all the available data up to the date we launch the prediction. We don't use data that has not been yet observed. For example, to predict January 2020 with lead 3 we only consider as training data until September 2019. As all the later dates contain dam states after the date we are trying to predict.

As it can be seen on step 2.c, we perform a reduction of predictors, as the number of predictors in some cases tend to be very large and we may overfit the model. For this we used the package Boruta_py which searches for features yielding the largest prediction skill. To give an illustration, if we consider the prediction of one of the sites, Salvajina, for lead 1 and experiment Obs+SeasPred (all observational data and Seasonal prediction indices as predictors), we have initially a total of 266 predictors considered. Although this is a very large number, current work is invested in reducing the number of initial predictors at the starting point. When the Boruta algorithm is used, the 266 predictors are reduced in the whole period of 1993-2016 to a mean of 41 covariates. In order to describe a bit further the distribution of these values: the 5th and 95th quantiles are 15 and almost 59, and the maximum and minimum number of covariates are 9 and 88. Hence the reduction of covariates used in the prediction reduces quite drastically in this method, reducing the chances of model overfitting.

Ordinary Least squares

This method consists of a standard Linear regression method (Ordinary Linear Regression, OLS) in which the covariate is the predicted precipitation in the region with lead 1-to-5 months (depending on the lead predicted), which is fitted to the state of the dam with lead 1-to-5 months. This approach is done monthly, which means that for January 2000 with lead 3, we fit the OLS of all available Aprils in the dam dataset with the precipitation predictions with lead 3 of the seasonal prediction system. Once the model is fitted, we predict April 2000.

Analog

This model takes advantage of the extensive record of dam observations provided by Celsia. It assumes that the present state of the dam will evolve in a similar manner to past situations. Hence, it provides as predictor of the dam evolution, the past evolution of the dam state starting in the closest month from which we start our prediction. For example, if we are predicting in January 2000 with lead 3, with the target of predicting April of 2000, the algorithm will look in the records for the December with the closest value to the last observed month, December

1999. Assuming the closest value is dated in December of 1978, we will follow the evolution of the dam during the winter 1978-1979 to obtain the targeted month of April of 1979. The value obtained in April 1979 will be used as our prediction of January 2000 with lead 3.

Prediction accuracy metrics

Three metrics considered to assess the goodness of the prediction over the period are R^2, Pearson correlation (R) and Root Mean Square Error (RMSE). The former metric, chosen initially, is complemented with the Pearson correlation to homogenize the validation of the results with other methods developed for Case Study 5. The latter metric is computed as percentage of the standard deviation of the observational anomalies of the dam.

2.3.3 Results

Leave-one-out Predictions

The performance of this algorithm is tested in the five sites provided by Celsia, although we focused initially in Salvajina. The resultant predictions for lead one (Figure 2.3.2) show as best models SeasPred, Obs+SeasPred and ObsOnly (All). The former two methods are expected to perform generally better, as they contain future predictions of climate indices from the seasonal prediction system, allowing the random forest to capture better the relationship of these indices with the dam state predicted. The model ObsOnly (All) is also ranked almost at the top of these models for lead 1, reflecting the ability of the random forest to capture relationships based on past observations. Nonetheless, as shown in Figure 2.3.3 these predictions decrease accuracy for increasing lead. From the poll of models, both Analog and OLS are the ones with worst performance in this approach, providing errors of the order of magnitude of one standard deviation and higher.

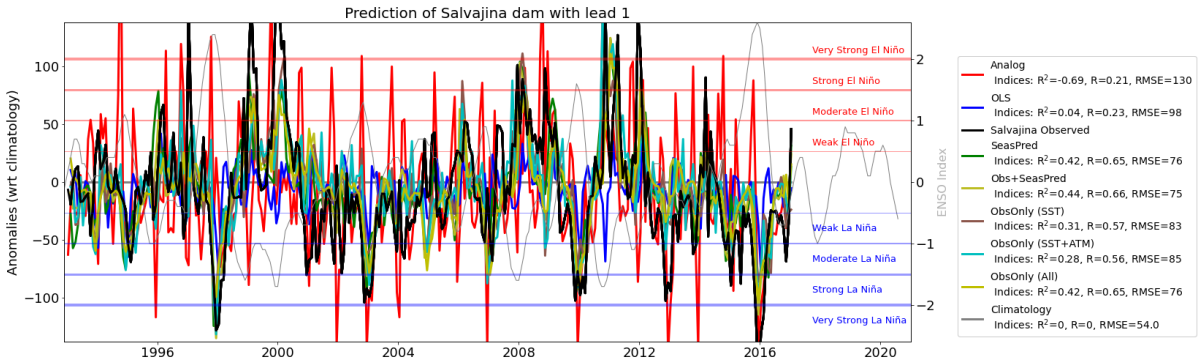

Figure 2.3.2 – Results of the predictions for all the models at Salvajina and lead of 1 month. The black line indicates the observations and the grey line the ENSO index. Horizontal lines indicate the strength of the ENSO index for periods of El Niño or La Niña.

When considering all the leads from 1-to-5, the results of the leave-one-out approach between 1993-2016 show as the best approaches Obs+SeasPred and SeasPred (Figure 2.3.3). Their RMSE of 78 and 76% of the standard deviation of the anomalies (solid lines in Figure 2.3.3), R^2 values for all the leads have a mean of 0.42 and 0.39 (dashed lines in Figure 2.3.3), and

correlation coefficients of 0.64 and 0.62 (dotted lines in Figure 2.3.3). These values are very similar for all the leads in the two models, with Obs+SeasPred having a higher spread in the scores. The performance of all the models using only past observations is very limited, decreasing as we increase the lead. In average the three models reduce their R^2 and R by 68% and 36% respectively and increase the RMSE by a 17%. Both OLS and Analog method provide poor prediction skill compared to all the other models.

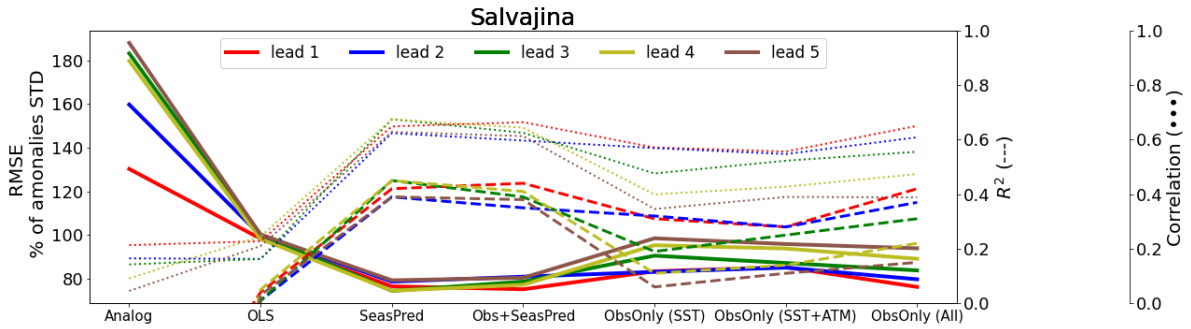

Figure 2.3.3 – Scores of the predictions at Salvajina for all the leads (colours) and all the methods (x-axis). RMSE as percentage of the standard deviation of the anomalies of the observations (solid lines). R^2 coefficient (dashed lines) and Pearson correlation coefficient (dotted lines).

The scores of the other four sites (Figure 2.3.4) depict a reduction of the prediction skill for all the methods. The methods performing better are still SeasPred and Obs+SeasPred, but the mean with respect to all the leads of their skills are reduced to values between 0.13-0.26 for the R^2, 0.38-0.51 for the correlation (Pearson correlation), and RMSE between 83-92% of the respective STD of each site anomalies. The rest of the methods underperform corresponding to the previous ones. As with Salvajina, the method based on all available past observations, ObsOnly (All), provides also improved prediction skills when compared to the climatology for low leads, lead 1 and 2.

In these values we did not account for the site Digua, which provides by far the worst results of all the five sites with no method providing any improvement of the prediction skill. We noticed that indeed this site was located downstream from Anchicaya, therefore making its evolution not completely dependent of climatological values and more related with the dam management from the energy company.

"Real-time" Predictions

The "real-time" predictions are based on using as training set only available past data. Its results at Salvajina provide a reduced skill compared to the previous approach (Figure 2.3.5). With scores for the best performing methods only obtaining positive R^2 for lead 1, correlations of less than 0.31 for all leads and RMSE below the 45% of the anomalous observational STD. Despite these results, the ultimate goal of this model was to provide better estimates than the ones considered by the energy company, and this was accomplished.

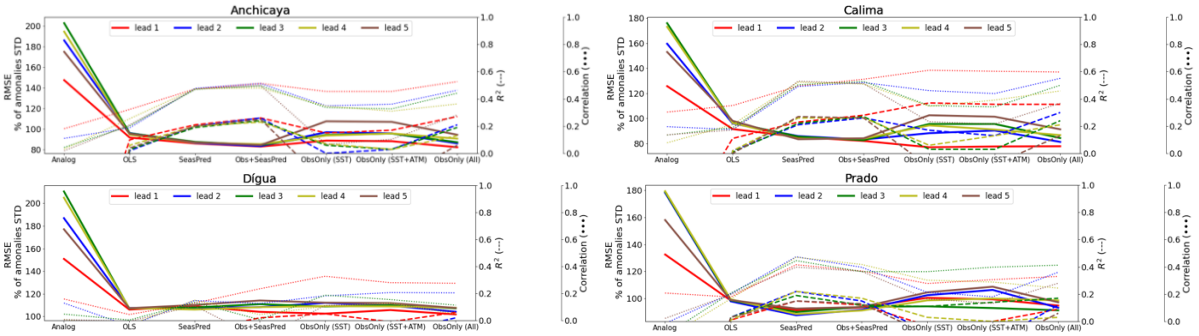

Figure 2.3.4 – Scores of the predictions at the other four sites of Celsia: Anchicaya, Calima, Dígua and Prado. RMSE as percentage of the standard deviation of the anomalies of the observations (solid lines). R^2 coefficient (dashed lines) and Pearson correlation coefficient (dotted lines).

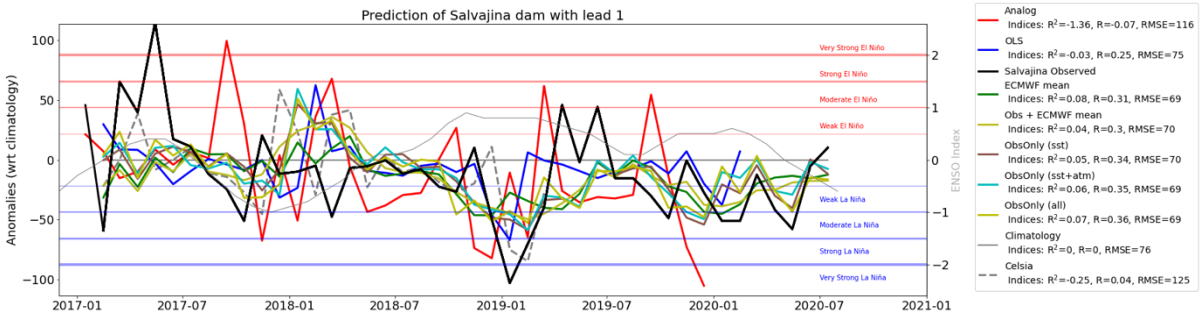

Figure 2.3.5 – Predictions of Salvajina with lead 1 for all the models in the period 2017-2020 performing "Real-time" predictions approach.

These results are not as good as expected based on the hindcast data, but further analysis should be done. Potential reasons behind the low performance could be simply a matter of the period that we are testing with no strong ENSO events limiting the predictability. The increased reliance on renewables for energy provision could have also an impact on the performance at this later period. An increase in the provision would increase the dam variability associated to an enhanced human induced management, which could decrease the direct relationship between climate variability and the dam variability. But further information of Celsia would be required to confirm this possibility.

After achieving the first goal of providing an improved estimate of the river flow for the end-user of Case study 5, this work is being developed further with the goal of a scientific paper. The work explores the added value and limits of seasonal prediction information used as predictors for variables not directly obtained by seasonal forecast systems (i.e. river flow) or poorly represented by these systems (i.e. precipitation).

2.4 Dynamical vs. statistical seasonal forecasts

Seasonal predictions of key atmospheric variables are an important area in climate science, because of its large value for a wide range of end users (e.g. Rodriguez et al., 2018; Demirel et al., 2015; Torralba et al., 2017). These seasonal forecasts can be produced by either a statistical empirical seasonal forecasting system or a dynamical forecasting system.

Statistical empirical methods have been used extensively (e.g. Barnston et al., 1999; Landsea and Knaff, 2000; Eden et al., 2015) and are based on observed relationships between certain predictors and the forecasted atmospheric variables (predictands). Dynamical forecasting systems on the other hand are based on numerical models that represent the governing processes in the atmosphere, ocean and land surface and their non-linear interactions. A disadvantage of the dynamical models is that they are inherently complex and computationally expensive. Furthermore, their model output often needs further calibration due to model drift towards their preferred climate state. Statistical models do not suffer from these two issues because their relationships are based on the observations themselves. Furthermore, the models tend to be relatively simple and easy to interpret. Arguably, statistical models can sometimes be too simple in order to capture the intricate non-linear relationships among predictor variables and the predictand.

Though already many comparison studies have been done between statistical and dynamical forecasting systems (e.g. Qian et al., 2020) and combining both methods to so-called hybrid models (e.g. Schepen et al., 2012; Zhang et al., 2016), these are often done at a regional scale. Here we analyze a suit of statistical models and assess their added information relative to a suite of dynamical models on a global scale.

We present an update of the relatively simple empirical statistical forecasting system from Eden et al. (2015) and implement more advanced statistical models based on tree-based regression systems. We assess the forecast skill of the statistical models and analyse their added value relative to dynamical seasonal forecasting models in a single and multi-model forecasting setup.

2.4.1 Models and Data

We will apply a suite of statistical methods for the forecasting system, ranging from relatively simple linear regression models to more advanced tree-based regression models. All models use the same predictors, however there are some differences in permutations of the predictors. We provide forecasts for near surface temperature (T2M) and precipitation.

2.4.1.1 Data

For T2M we use the GHCN-CAMS dataset (Fan and van den Dool, 2007) over land and ERSSTV5 (Huang et al., 2017) dataset over sea. For precipitation we use the GPCC dataset (Schneider et al., 2014). The predictors we use can be divided into local predictors (spatio-temporal data), large scale climate indices and the CO2 equivalent forcing (CO2EQ) as a

predictor for the long-term trend. For the local predictors we use persistence (for T2M and Precipitation) and cumulative precipitation (for T2M). For the large-scale climate indices, we use the El-Niño Southern Oscillation (ENSO), the pacific decadal oscillation (PDO), Atlantic multidecadal oscillation (AMO), the Indian ocean dipole (IOD) and the North Atlantic Oscillation (NAO). Though there are many more teleconnections active throughout the ocean and atmosphere, we have selected these because of their predictive power on lead times in the order of months. We quantify ENSO through multiple indices, namely the NINO34 and NINO12 index and the warm water volume (WWV). All SST based indices are calculated on the ERSST V5. For T2M we use the GHCN-CAMS dataset (ref) over land and ERSST v5. Table 2.4.1 gives an overview of all the predictors used and which dataset is used. All data is re-gridded to a 1x1 grid and we use the time period ranging from 1961 to current.

Table 2.4.1: Overview of predictand and predictor data

Predictands	
T2M	GHCN-CAMS
Precipitation	GPCC
Predictors	
CO2EQ	
NINO34	ERSST V5
NINO12	ERSST V5
WWV	POAMA / PEODAS
PDO	ERSST V5
AMO	ERSST V5
IOD	ERSST V5
Precipitation	GPCC
Persistence	-

2.4.1.2 Statistical Empirical Models

All models are set up in the same way but differ mainly in the predictor selection and fitting routine. All the models set up aim to predict the next 3-month mean, i.e. a forecast initiated early January will produce a forecast for FMA. Models are fitted for each grid point individually. The predictor and predictand data are first detrended based on a linear regression with CO2EQ, in order to have a stationary dataset. Then, if necessary, a predictor selection routine is done to assess the relevant predictors. Next, the model is fit to create a forecast, and hindcasts from 1961 to current are created using Leave-1year-out cross-validation. The trend previously subtracted is then added to the model in order to get the final results. All models have a global coverage and a horizontal resolution of 1 degree.

2.4.1.3 Multiple Linear Regression (MLR)

The MLR model can be seen as an updated version of the relatively simple empirical statistical forecasting system from Eden et al., 2015. It forecasts the next 3-month average based on the previous 3-month mean predictors. As an example, a forecast of T2M issued in January is based on OND predictor data and predicts FMA.

A recent update, relative to Eden et al. (2015), to this method is to also include 2nd order information in the predictors, by including the 3-month trend as predictor. In order to avoid overfitting by introducing too many predictors, for the climate indices an intermediate MLR model is used to predict the future state of the predictor, based on its previous 3-month mean and 3-month trend. The future state of the predictor is then used to fit the actual MLR model. For the NINO34 index this greatly reduces the spring predictability barrier (Figure 2.4.1). For persistence, an extra persistence trend predictor is added. Figure 2.4.1 shows the added value of using the 3-month trend as a predictor for the different climate indices.

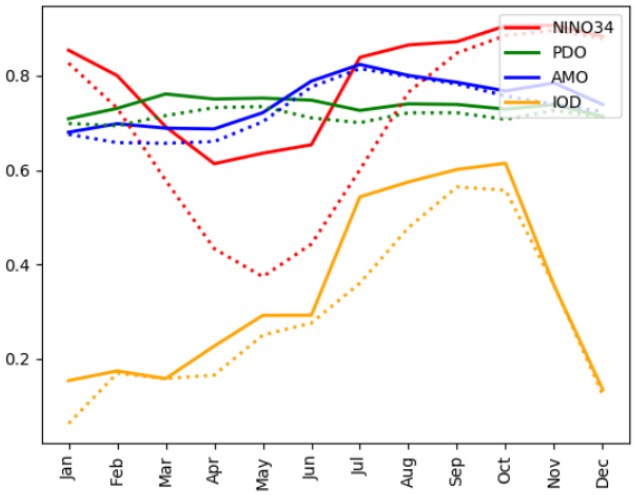

Figure 2.4.1: Correlation between the climate indices. The dashed lines represent the lagged correlation between the 3-month mean and the 3-month mean 3 months ahead. E.g., January shows the correlation between OND average with the next FMA. The solid lines show the correlation between the forecasted climate index and the observed climate index. The forecasted index is based on a MLR model with the 3-month mean and 3-month trend as predictors.

Another update to the original forecasting system is to have a stricter predictor selection routine. Previously, only predictors were added that have a significant ($p<0.05$) correlation with the predictand. Though this removes non-relevant predictors, it does not account for co-variability between predictors. This can e.g. be an issue in the NINO34 region, between the NINO34 index and persistence of SST. To overcome this, we first compute the correlation between the potential predictors and predictand. The predictor with the highest significant correlation is then chosen as a predictor, after which the linear relation between the chosen predictor and predictand is removed from the predictand time series. Hereafter, we again compute the correlation between the remaining potential predictors and the residual predictand time series, and include the strongest and significant predictor in the regression model. This is continued until there is no significant relation between a potential predictor and predictand, and is done for every grid point individually.

The ensemble is calculated by randomly sampling from the residuals (forecast error) of the model fit. If there is poor predictability, the errors will be large thus the ensemble spread

relatively large. If there is good predictability, the errors will be small and thus the ensemble spread relatively small.

The advantages of using the relatively simple MLR method is that we can identify the individual contribution of each predictor and works well on relatively small sample sizes. The disadvantages of using MLR is that it assumes a normal distribution and homoscedasticity, is sensitive to outliers and assumes a linear relationship.

Note that besides MLR we also tested Lasso and Ridge regression, but found no improvement relative to MLR hence we did not proceed with these methods.

2.4.1.4 Random Forest Regression (RFR)

Random forest regression is a tree-based ensemble regression model, which is a popular machine learning tool used for many different forecasting problems and research fields. Random forests are constructed by individual decision trees. Decision trees can make very accurate predictions on the data it was trained on. However, it generally leads to very bad results on new (testing) data. To circumvent this, trees can be built using a bootstrapped sample of the original training data. The end result is achieved by taking the average of all these trees. This method is known as bagging, i.e. taking the aggregate of all the different trees based on bootstrapped training data. The main advantage of this method is that it performs much better on testing data, thus leading to better generalization.

With these ML methods, there are multiple parameters that have to be defined. These are the maximum depth of the tree (max_depth), number of trees in the forest (n_estimators, i.e., how many trees in the forest), minimum samples per leaf (min_samples_split) and the maximum features (predictors) to choose from while making the tree (max_features). Given that the model is fit for each grid point, the optimal settings will differ per grid point. A first analysis showed that the RFR model was most sensitive to the maximum depth, hence we chose to perform a parameter selection routine on the maximum depth ranging from 1 to 7. Table 2.4.2 lists the parameter settings. The parameter combination with the lowest mean square error (MSE) is used for the final model. Generally, in regions with low forecast skill the maximum depth is kept smaller to avoid overfitting, whereas in regions with larger predictability the maximum depth is higher.

Table 2.4.2: Parameter combinations RFR models

Parameter	Setting
Nr. of estimators	50
Max. depth	[1 .. 7]
Max. features	3
Min. samples split	?

RFR models tend to be ‘data hungry’ (Van der Ploeg et al., 2014), i.e. require a large training sample in order to acquire stable models. The data used in this study covers 1961 to current, thus ~60 samples if the model is fit for each month individually and ~700 samples if the model is fit using all months together. Preferably the model is fitted per month individually because

predictor-predictand relations can differ strongly for the different months. However, initial tests pointed to stronger overfitting using only 60 samples relative to using the full dataset. This is why we constructed two RFR models, one that fits a model for each month individually (RFR-M), and one model that uses the full sample (all months together, RFR-Y). Note that in the RFR-Y model we divide the predictand data by its monthly standard deviation prior to the model fit, and multiply the forecast again by its monthly standard deviation. This is done because variability can change over the different months, especially for precipitation, which can lead to unrealistic high variability over months with relatively low variability.

We increased the number of predictors in the model by adding more permutations of the predictor data to the model. Besides the 3-month mean and 3-month trend, as done in the MLR model, we also added the 5-month mean, and 1-month values at lags of 1, 3 and 5 months. We tested multiple predictor selection routines, but found no clear reduction in forecast error and greatly increased the model run time. Hence, we do not use a predictor selection routine for the RFR models.

Probabilistic forecasts are constructed by considering all trees of the random forest, and not only the average or median value over these trees as generally done. Given that our nr. of estimators (trees in the forest) are 50, we have an ensemble size of 50. The advantages of RFR relative to MLR is that it assumes no distribution, and can handle non-linear relationships and heteroscedasticity.

We also tested several other tree-based regression models such as the gradient boosting method and regression enhanced random forest, but these methods showed no clear improvement of standard RFR hence we did not proceed with these methods.

2.4.1.5 Dynamical models

In order to assess the added value of statistical models, we compare the statistical models to a suite of dynamical seasonal forecasts (listed in Table 2.1.1 in Section 2.1.1). The seasonal forecasts are all bias corrected on a monthly basis.

2.4.2 Verification

Despite the large improvements over recent times in observational products, there still remains relatively large observational uncertainty for certain regions. Different observational products can differ considerably making it non-trivial to select a single observational product for verification. Figure 2.4.2 shows the observational uncertainty, quantified by the disagreement (standard deviation) between multiple observational and reanalysis products (listed in Table 2.4.3), Here we first compute the standard deviation (std) per time step between the different observational products, and then average this std over all months and years. It is no surprise that generally speaking there is a larger uncertainty around regions with less observations. However, there are many regions with a standard deviation around 0.5 °C, indicating a spread of around 2 °C. The statistical empirical models are all biased towards their own reference product, whilst the dynamical models are also all biased towards their respective observational product used for their initialization. Hence, which model performs better will depend strongly on which observational product is used as reference (or truth).

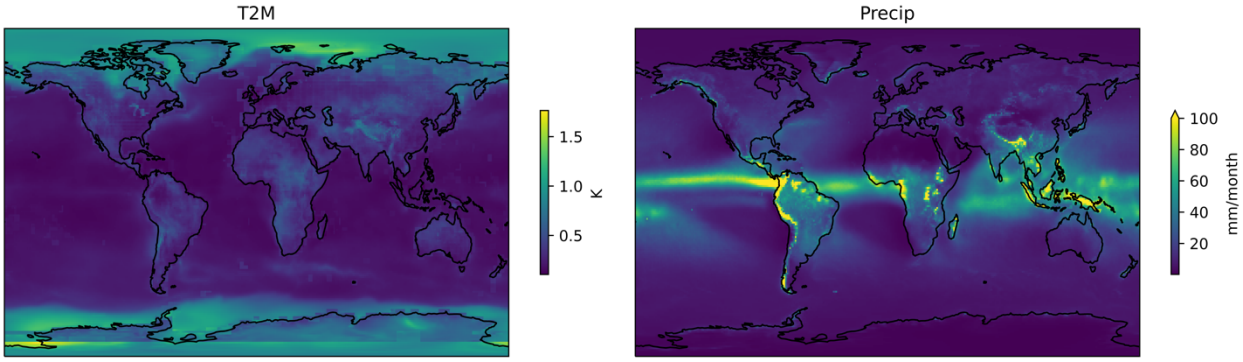

Figure 2.4.2: Observational uncertainty of T2M and precipitation quantified by the average standard deviation between all products used, by first computing the std between all products and then averaging over all the full time period.

To circumvent this potential bias, we use the average of multiple observational and reanalysis products (ENS_OBS) as a reference dataset to evaluate the forecast skill.

Table 2.4.3: List of observational and reanalysis products used.

Product	T2M	Precipitation	Reference
ERA5	x	x	Copernicus Climate Change Service, 2017
JRA-55	x	x	Kobayashi et al., 2014
MERRA-2	x	x	Gelaro et al., 2017
ERSST V5	x		Huang et al., 2017
CRUTEM	x		Harris et at., 2013
GPCC		x	Schneider et al. 2015
GISTEMP	x		Lenssen et al., 2019

2.4.3 Results

The forecast skill is quantified by the continuous ranked probability score (CRPS), which is an often-used measure for probabilistic forecasts. It quantifies the error based on the quadratic measure of the difference between the forecast cumulative density function and the observed value. We use the CRPS skill score variant (CRPSS), by directly comparing the CRPS of the forecast to a reference forecast (CRPSS = 1 - [$CRPS_{for}$-$CRPS_{ref}$]). Positive values indicate the forecast outperforms a reference forecast.

2.4.3.1 MLR

First, we will look at the forecast skill of the relatively simple MLR model. Figure 2.4.3 shows the CRPSS of T2M and PRECIP for the JJA and DJF forecasts, initiated in respectively May and November. We use a climatological forecast as reference, which is constructed by randomly sampling 51 values from the climatology with leave-1-out cross-validation.

It is clear from the T2M figures that forecast skill is greatest in the tropical regions, owing to the large influence of ENSO. Also, skill over the ocean is generally larger because of the stronger persistence of anomalies. In JJA we also find some skill over Europe, which is mainly related to the long-term trend.

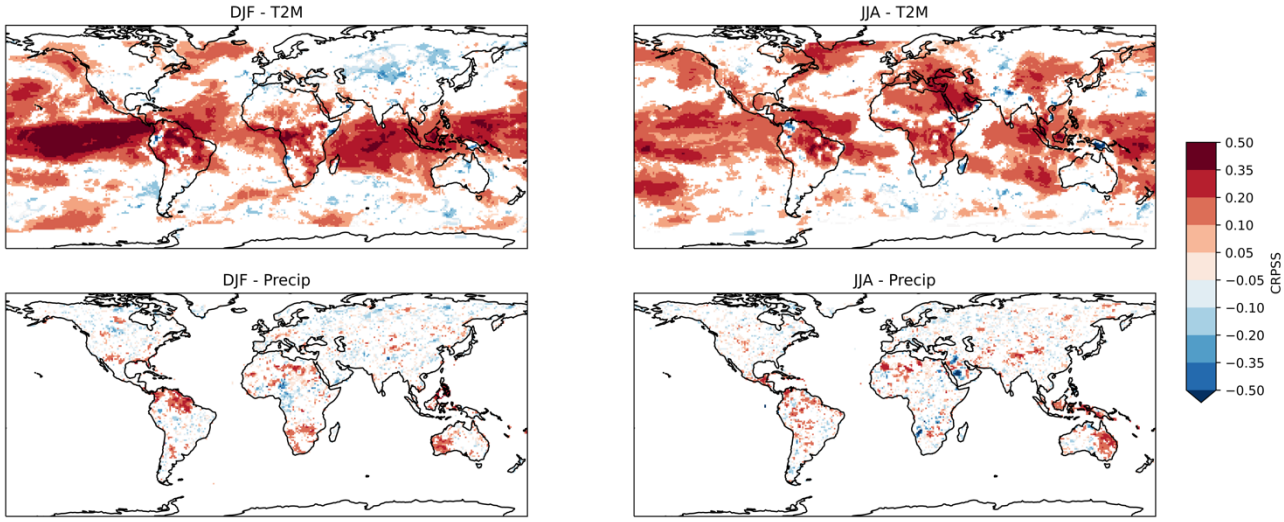

Figure 2.4.3: Forecast skill (CRPSS) with a climatological forecast as reference for the MLR model, for T2M and Precip and for DJF and JJA. The skill score is based on data from 1980 to 2016.

For precipitation, the forecast skill is generally much lower. The teleconnections with ENSO do provide some skill over the northern part of South America, Australia, and southern part of North America.

Negative values (worse than a climatological forecast) for both T2M and precipitation are to some extent caused by overfitting, but mostly due to differences in between GHCN (used as the observational estimate in the model fit) and ENS_OBS.

Note than an extensive evaluation of the added benefit of the individual predictors is already done in Eden et al. (2015). In order to further understand and study the sources of predictability, we have made an online interactive application which allows analyzing the forecasts and the sources of predictability in more detail (*http://climexp.knmi.nl/kprep_fc*).

2.4.3.2 RFR

Next, we assess the forecast skill of the RFR-M model, where the MLR forecast is used as reference (Figure 2.4.4). Hence, red values indicate the RFR-M model outperform the MLR model, whereas blue values indicate the MLR model outperforms the RFR-M model. The RFR-M outperforms the MLR forecasts for several regions, but there are also quite some regions where MLR outperforms the RFR-M model. There are distinct regions where one model performs better than the other, such as Russia for MLR and the Western Pacific for RFR-M. For precipitation, MLR mostly outperforms the RFR-M model with the exception of an area in Southern America for DJF and Northern Africa in JJA.

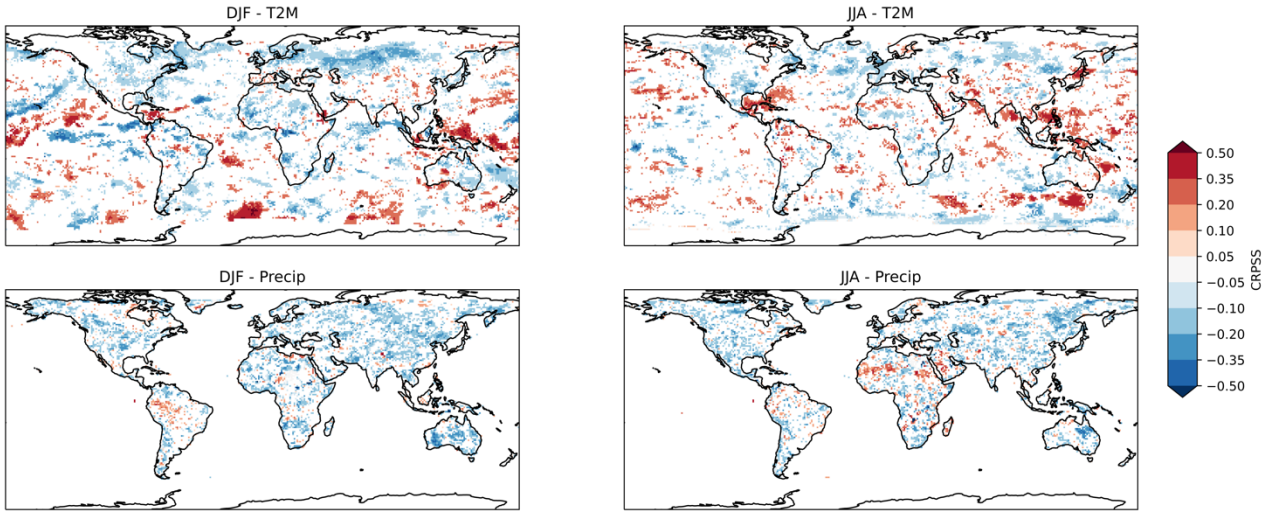

Figure 2.4.4: CRPSS of RFR-M with MLR as reference forecast, for T2M and precipitation on the forecasts valid for winter (DJF) and summer (JJA). The skill score is based on data from 1980 to 2016.

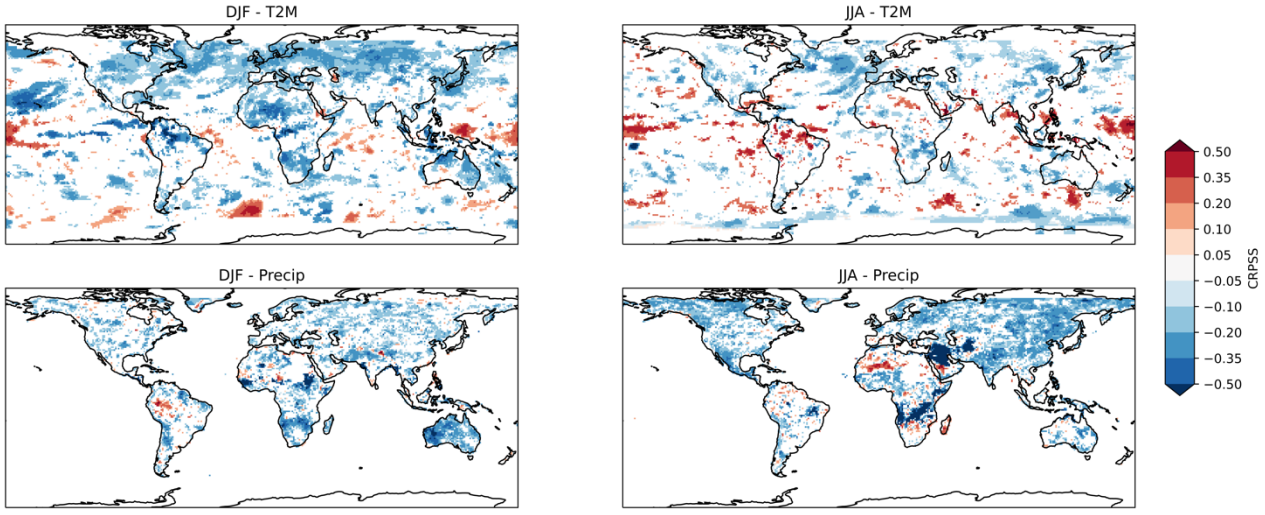

Figure 2.4.5: Same as Figure 2.4.4, but now for RFR-Y instead of RFR-M.

When comparing RFR-Y with MLR (Figure 2.4.5), we again find some regions where RFR-Y outperforms MLR, however there are considerably more regions where MLR outperforms RFR-Y. The RFR-Y model fails to reproduce some of the interannual variability (not shown) in the northern regions, where MLR and RFR-M are capable of reproducing some of the variability. It seems that for a lot of regions the seasonal relation between predictor and predictand differs considerably, leading to worse results when pooling all months together. For precipitation the MLR model seems to outperform RFR-Y in almost all regions. Again, also for precipitation the RFR-Y fails to reproduce the interannual variability in most regions.

The results indicate that using more advanced models does not necessarily lead to better results. RFR models generally need a large training set in order to create stable models, and it seems that ~60 years of data is not a large enough sample to really outperform MLR. By pooling all months together (RFR-Y) the sample size is largely increased (factor 12), but at the cost of losing the individual predictor-predictand relations on a monthly basis. Especially for ENSO, which is phase locked to the seasonal cycle, the relations differ strongly throughout the year, making RFR-Y less skilful than RFR-M.

2.4.3.3 Statistical vs dynamical

The advantage of statistical models relative to dynamical models is the low computation costs and an easier understanding of the sources of predictability, either through the regression coefficients (MLR) or feature importance's (RFR). It is however important to know whether statistical models provide added information relative to dynamical models. All analysis in this section is performed on data ranging from 1994 to 2016.

In order to assess the added value of the statistical empirical models, we compare them to a set of dynamical forecasts (listed in table 2.1.1). We calculated the CRPS of each model (ENS_OBS as observational estimate) and selected the best performing model (lowest CRPS) per grid point (Figure 2.4.6). The labels in Figure 2.4.6 denote the type of model which performed best. It is clear that for most regions one of the dynamical models performed best. However, especially in the JJA, both for T2M and precipitation forecasts there are still quite some regions where either a RFR model or MLR performed best.

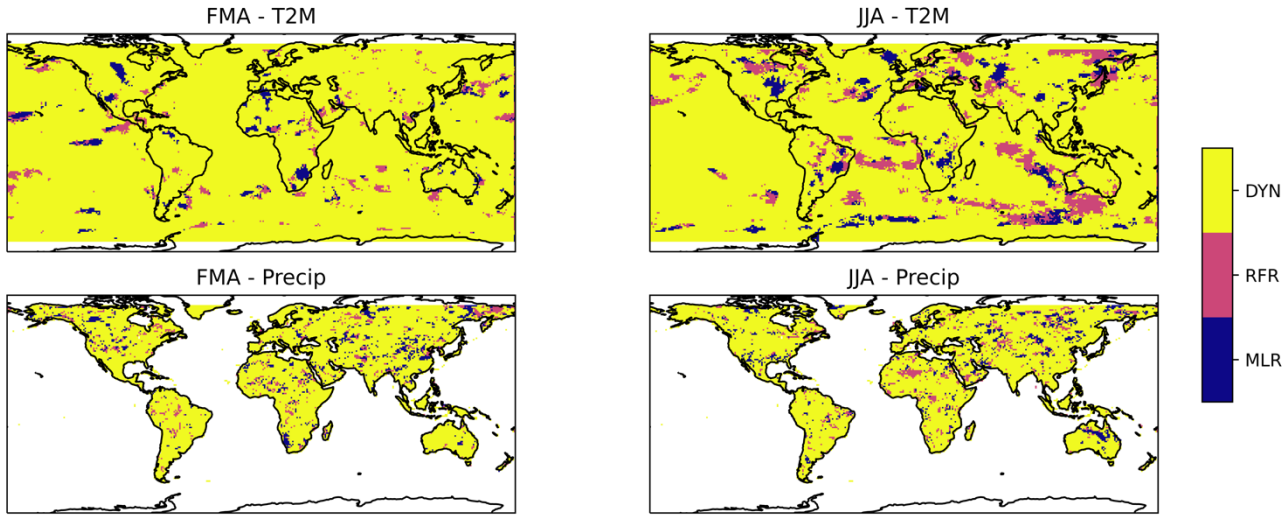

Figure 2.4.6: Best scoring model type per grid point, based on minimum CRPS.

2.4.3.4 Added value in a multi-model framework

Though the results in Figure 2.4.6 indicate that mainly dynamical models provide the best individual forecast, this does not automatically indicate that there is no added value in using statistical forecasts. In general, a multi-model combination of seasonal forecasts tends to

outperform single seasonal forecasts (Hagedorn et al., 2005, see also previous sections in this report). Hence, in order to fully assess the added value of statistical models, we constructed a multi-model forecast of 5 models with every possible combination with the dynamical and statistical models available. This number was chosen as the average number of models in the optimal multi model combination was estimated to 4.5 models for the two variables over the domains investigated by UL in their study is found from Table 2.1.4. The model with the lowest RMSE was selected, after which we calculated how many statistical models were in this specific model. This analysis was performed for each grid point individually. The results can be found in Figure 2.4.7. It is clear that relative to Figure 2.4.6 there are many more regions where a statistical model provided additional information in a multi-model framework. For JJA and T2M there are even quite some regions (North America, South America, Europe) where at least 1 of the three models are statistical models. For precipitation, the results are a bit more scattered but do point to the same conclusion that in a multi-model framework there is regionally added value by combining statistical with dynamical models.

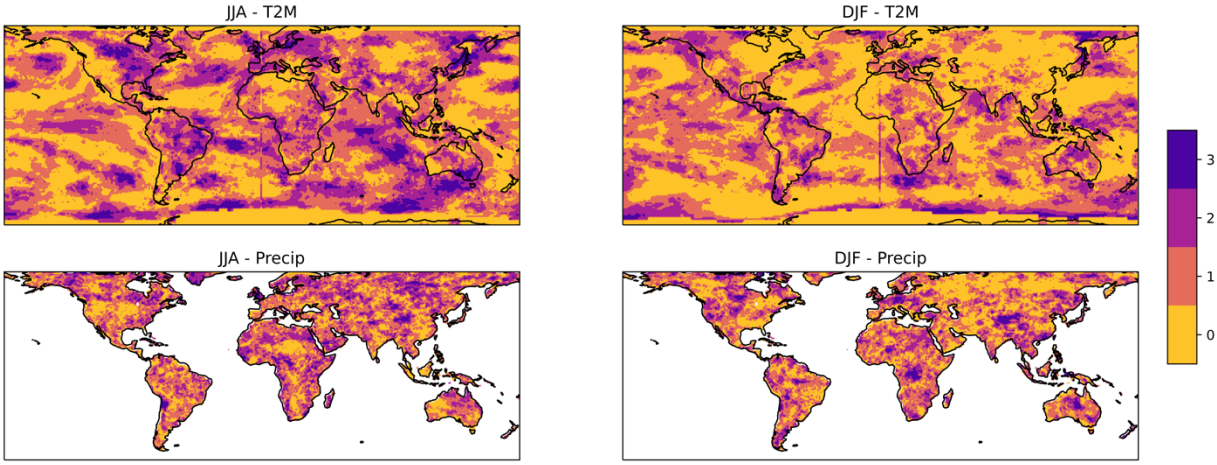

Figure 2.4.7: Added value of statistical models in a multi-model framework. The values indicate how many statistical models were listed in the best performing multi-model combination.

In order to assess which models performed best, we also computed where a certain model was selected in the best model combination. These results can be globally aggregated to form a global coverage percentage (Figure 2.4.8). From these results we can see that C3S-ECMWF (ECMWF-S5) is the best performing seasonal forecasting model, both for T2M and precipitation. The statistical models do score lower than the dynamical models, especially in DJF. For precipitation, the differences between the models are smaller. This is most likely related to the lower overall forecast skill, which will make the best model selection more random.

Note that the analysis shown in Figure 2.4.7 and 2.4.8 should be treated as a qualitative analysis as no significance testing is performed.

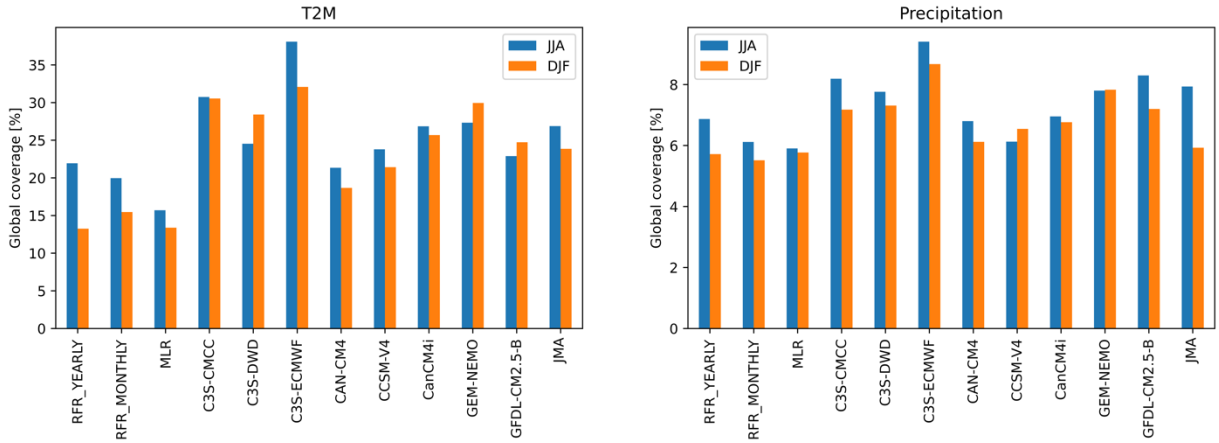

Figure 2.4.8: Global coverage of models present in the best model combination. quantified by the global coverage.

2.4.4 Conclusions

In this analysis we analysed the forecast skill of relatively simple and more advanced statistical empirical forecasting systems, and assessed their added value relative to dynamical seasonal forecasting systems. The relatively simple seasonal forecast based on multiple linear regression performs quite well. It has good skill in the tropical regions, where there are strong teleconnections with large scale climate indices such as ENSO. Persistence of anomalies and the long-term trend is also a large source of forecast skill.

RFR models improve the forecasts locally, but the small sample size hampers their forecast skill. By using the full sample (all months pooled together) the forecasts become more stable (less overfitting). This improves the forecast in certain regions, but mostly reduces forecast skill in other regions because it loses the individual predictand-predictor relations that differ throughout the year. Hence, we find no 'best' model for all grid points, but more an ensemble of statistical empirical models whose skill depends on the region considered.

When comparing the statistical models to a suite of dynamical models, we find that in general the best individual models are one of the dynamical models, though the specific model varies depends on the area. There are some regions where the best forecast skill is obtained by a statistical model, but this is rather limited. In a multi-model framework, there are numerous regions where the multi-model average forecasts are improved by a combination of statistical and dynamical models instead of only using a combination of dynamical models. For JJA temperature e.g., there is added value in large parts of Northern and Southern America and Europe. Given that marginal improvements in seasonal forecasts of precipitation and temperature are already very useful for the energy sector, these results highlight the need for adding statistical models to multi-model ensemble seasonal forecasts.

A scientific paper is in preparation describing the methodology of the statistical seasonal forecasts and its added value relative to dynamical seasonal forecasts.

2.5 Signal inflation in Seasonal climate predictions

2.5.1 Can inflation of the forecast signal improve the seasonal climate predictions over Europe?

In the North Atlantic and European region climate forecasts are known to have too small a signal to noise ratio in the models compared to observations (Eade et al 2014, Scaife and Smith 2018). Smith et al (2020) found that when the underprediction of the signal was accounted for, skilful winter predictions of decadal mean climate around the North Atlantic basin could be obtained. This result relied on the use of a very large multi-model ensemble, a skilful forecast of the North Atlantic Oscillation (NAO) Index and the sub-selection of ensemble members with a realistic NAO magnitude. Here we assess whether this approach can be used to improve seasonal climate forecasts over the European and North Atlantic region.

2.5.2 Data and methods

The skill in forecasting winter (December, January, and February) mean temperature, precipitation, and mean sea level pressure (MSLP) is assessed from a November start date, giving a 1-month lead time. The observational datasets used to assess forecast skill are listed in Table 2.5.1. Seasonal forecasts have been collated from modelling centres across Europe and North America (see Table 2.5.2). Retrospective forecasts, known as 'hindcasts', are available over a 24-year period, from 1993-2016. Different models have different climatological biases. To take this into account, for a given field, each ensemble member forecast is represented as an anomaly relative to its own model's climatological average. After this step, all ensemble members are treated as though coming from the same model. An ensemble mean forecast is created by averaging across the individual member's seasonal anomalies.

Table 2.5.1: The observational datasets used in the seasonal forecast assessment

Variable	Observational dataset
Mean sea level pressure (hPa)	Hadley Centre sea level pressure (HadSLP2r)
2m air temperature (K)	HADCRUT4
Precipitation (mm/day)	Combined precipitation dataset v2.3 (GPCP)

Table 2.5.2: The models used in the seasonal forecast assessment

Centre Name	Model version	N. of ens. members
Met Office, UK	HadGEM3 GC2.0 (C3S v14)	28
Météo-France	7 (C3S v7)	25
CMCC, Italy	CM2 (C3S v3)	40
ECMWF	SEAS5 (C3S v5)	25
DWD, Germany	GCSDv2.1 (C3S v2)	30
NCEP, US	CFSv2 (C3S v2)	24
JMA, Japan	MRI-CPS2 (C3S v2)	10
ECCC, Canada	CanSIPSv2	20
	Total	**202**

An assessment of forecast skill is made by calculating both the Pearson correlation coefficient and the root mean square error (RMSE) between the observed climate and two sets of forecasts.

1) The first forecast set is simply the 'raw' ensemble mean across all available ensemble members (all 202 members). A raw ensemble mean forecast is calculated for MSLP, Temperature, precipitation and the NAO and referred to with a subscript 'ens'.

2) The second forecast set is an ensemble mean across a limited selection of forecast members, following the Smith et al (2020) methodology to inflate the forecast signal. This two-step process is described below:

Step 1: Inflation of the signal strength using the full ensemble.
The raw ensemble mean NAO forecast (NAO_{ens}) is inflated by multiplying by the Ratio of Predictable Signals (RPS), as defined below:

$$NAO_{inf} = NAO_{ens} * RPS \qquad \text{(Eq. 2.5.1)}$$

$$RPS = RPC \frac{\sigma^o{}_{total}}{\sigma^f{}_{total}} \qquad \text{(Eq. 2.5.2)}$$

With $\sigma^o{}_{total}$ the observed standard deviation of the NAO , $\sigma^f{}_{total}$ the standard deviation of all the ensemble members NAO index, and RPC the Ratio of Predictable Components. The latter is defined as

$$RPC = ACC / \left(\frac{\sigma^f{}_{sig}}{\sigma^f{}_{total}}\right) \qquad \text{(Eq. 2.5.3)}$$

The ACC is the anomaly correlation coefficient between observed and model ensemble mean NAO and $\sigma^f{}_{sig}$ the standard deviation of the model ensemble mean NAO.

Step 2: sub-selection of ensemble members to create the new forecast.
For a given winter, the 40 individual ensemble members which have the closest NAO value to that winter's inflated ensemble mean NAO value (NAO_{inf}) are selected from the 202 available. The forecasts of MSLP, temperature and rainfall for these 40 members are then averaged to give a new 'post processed' forecast. As the inflated ensemble mean NAO value changes each winter, the 40 members chosen will also vary between winters.

2.5.3 Results: Raw model skill

The raw skill in forecasting winter MSLP, 2m temperature and precipitation, using the full ensemble, is shown in Figures 2.5.1-3. Forecast skill is considered robust where the Pearson

correlation is statistically significant at the 5% level, using a 1-sided Fisher Z test (shown by stippling). Considering the North Atlantic and European region, skilful forecasts of MSLP are found in the Atlantic basin north of Iceland. Skilful temperature forecasts are found over the North Atlantic and parts of Scandinavia, whilst there are no regions of significant precipitation skill in Western Europe.

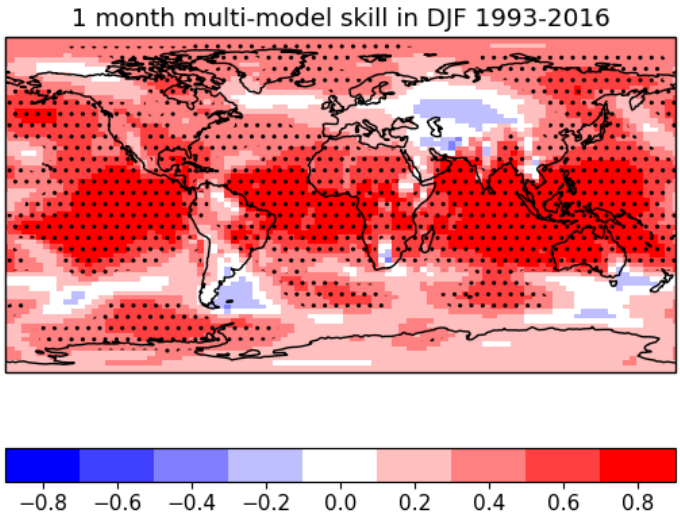

Figure 2.5.1: Raw winter MSLP skill. The Pearson correlation coefficient between observed and multi-model ensemble mean. Statistically significant skill at the 5% level is shown by stippling, using a 1-sided Fisher Z test.

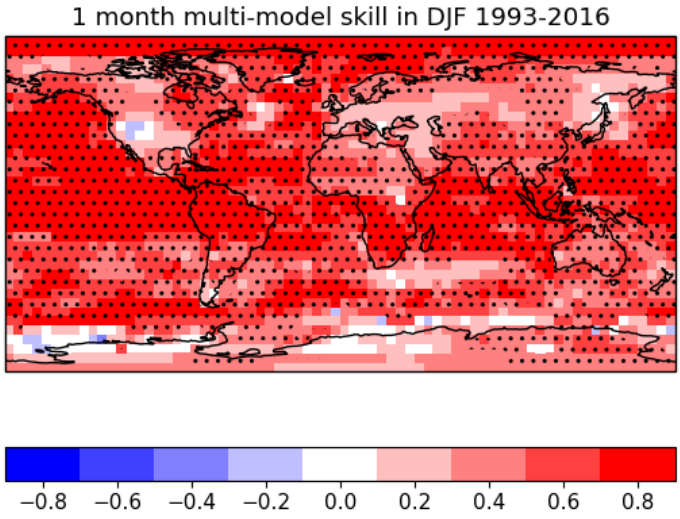

Figure 2.5.2: Raw winter temperature skill. The Pearson correlation coefficient between observed and multi-model ensemble mean. Statistically significant skill at the 5% level is shown by stippling, using a 1-sided Fisher Z test.

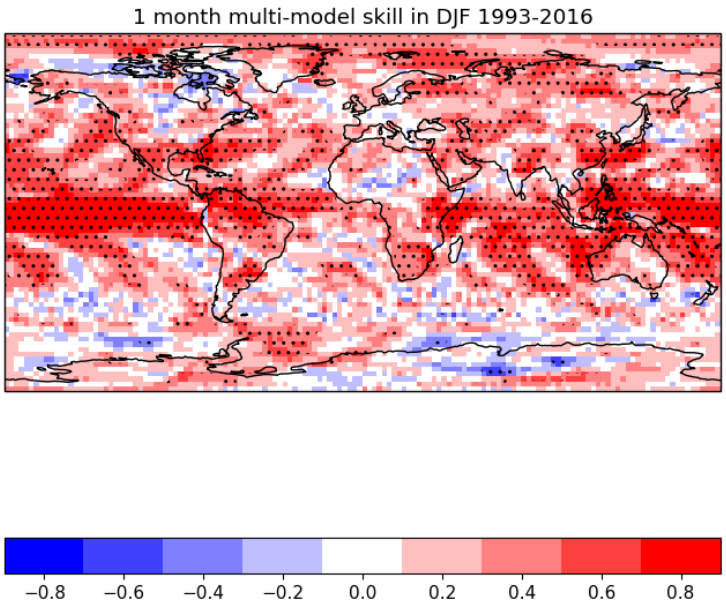

Figure 2.5.3: Raw winter precipitation skill. The Pearson correlation coefficient between observed and multi-model ensemble mean. Statistically significant skill at the 5% level is shown by stippling, using a 1-sided Fisher Z test.

The raw forecast of the North Atlantic Oscillation ($NAO_{ens)\ is}$ skilful, with a Pearson correlation of 0.48, as shown in Figure 2.5.4. As detailed above, NAO_{ens} is calculated using all 202 ensemble members. It had been hoped that this large ensemble would lead to a significantly more skilful forecast than that possible with individual models. This is not the case with the most skilful models having similar prediction skill to the ensemble mean, for example the UK Met Office model has a correlation of 0.49 and the German model a correlation of 0.47 (see Table 2.5.3).

Table 2.5.3: Winter mean NAO forecast skill, given by the Pearson correlation coefficient for each model individually and for the full ensemble mean. Statistically significant skill at the 5% level is shown by a star (*), using a 1-sided Fisher Z test.

Centre Name	NAO forecast skill (Pearson correlation)
Met Office, UK	0.49*
Météo-France	0.19
CMCC, Italy	0.28
ECMWF	0.32
DWD, Germany	0.47*
NCEP, US	0.29
JMA, Japan	0.22
ECCC, Canada	0.45*
Ensemble mean (NAO_{ens})	**0.48***

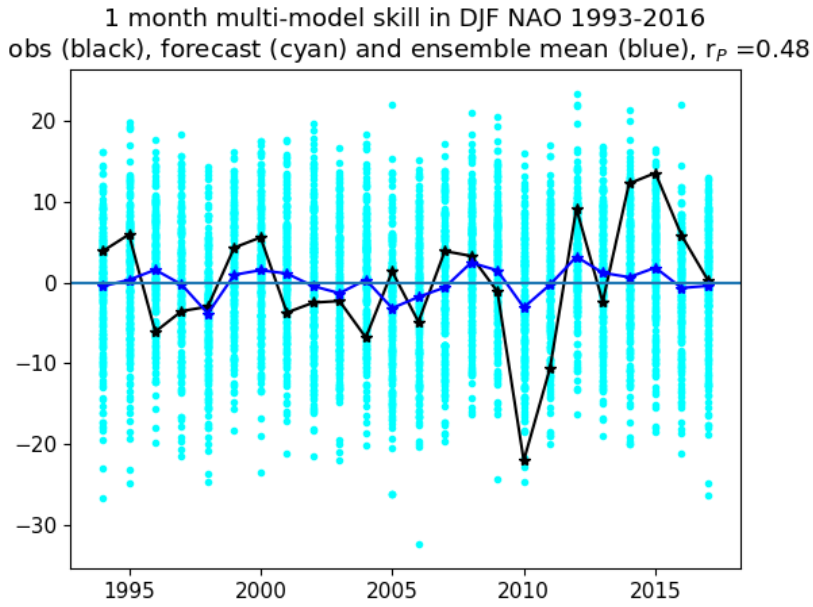

Figure 2.5.4: Winter mean North Atlantic Oscillation Index of observations (black), individual ensemble members (cyan) and multi-model ensemble mean (blue). The Pearson correlation coefficient (r_p) is given in the title.

2.5.4 *Post-processed forecast skill, the impact of strengthening of the forecast signal*

The Ratio of Predictable (RPS) signals for the NAO is 2.1 when calculated following equation 2.5.2. Consequently, the inflated NAO value (NAO_{inf}) for a given winter, is approximately two times larger than the original raw ensemble mean value (NAO_{ens}), following equation 2.5.1. The inflated ensemble mean NAO index (NAO_{inf}) is shown in red in Figure 2.5.5. This inflation process does not modify the correlation with the observed NAO.

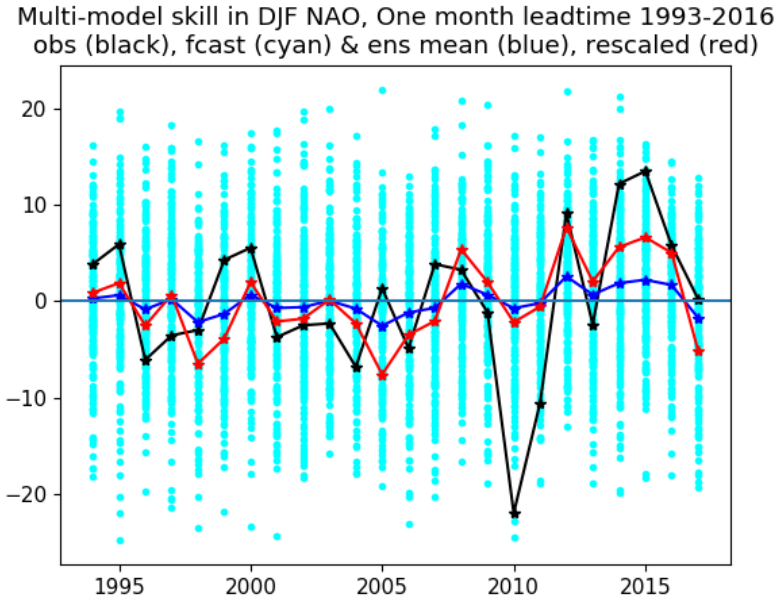

Figure 2.5.5: The impact of inflation: Winter mean North Atlantic Oscillation Index of observations (black), individual ensemble members (cyan) and ensemble mean (NAO_{ens}, blue) and inflated ensemble mean (NAO_{inf}, red).

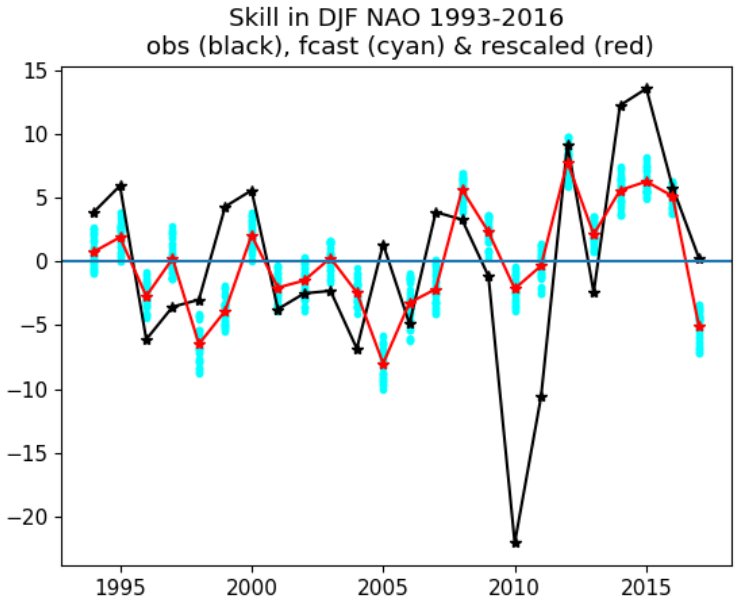

Figure 2.5.6: This figure is the same as 2.5.5, except that only the ensemble members chosen as part of step 2 are now shown in cyan. Winter mean North Atlantic Oscillation Index of observations (black), individual ensemble members (cyan) and inflated ensemble mean (NAO_{inf}, red).

Figure 2.5.6 shows in cyan the ensemble members that have been chosen in step 2. The 40 nearest ensemble members to the ensemble mean NAO value (in red) for each year have been selected. The corresponding fields of MSLP, temperature and precipitation for each of the selected members are then averaged to give the new 'post-processed' forecast.

Figures 2.5.7-9 show the change in root mean square error (RMSE) of the forecasts after post-processing, for MSLP, temperature and precipitation respectively. The forecast RMSE of MSLP increases over many areas of the North Atlantic after post processing, whilst over mainland Europe, there is little change. The post processed temperature forecast RMSE improves over eastern Europe/Russia, with little change over Western Europe. Over Europe the post processing method does not improve the rainfall forecasts. A similar picture is seen when using the Pearson correlation coefficient as the skill measure (not shown).

In conclusion, strengthening the forecast signal does not give significant benefits for seasonal prediction of winter mean temperature and rainfall over Europe. This is in contrast with the improvements in decadal prediction of surface climate found in Smith et al (2020). The higher prediction skill in the NAO at the decadal timescale (r =0.79 compared to r=0.48) is likely the main reason for the difference, driven in part by the much larger ensemble available. The inflation factor for the NAO signal is also much larger in the decadal prediction setting, with an RPC of 11 compared to 2 for seasonal prediction.

The strengthening of the forecast signal will consequently be having a much larger impact on the decadal predictions than for the seasonal predictions. In addition, over the decadal hindcast period, the climate change signal is much larger than that over the shorter seasonal hindcast period. The signal inflation method may therefore be better balancing the climate change and dynamical influences in the decadal predictions, giving the improvements seen. If additional seasonal forecasts become available and the ensemble mean NAO skill improves, it would be worth revisiting whether the signal inflation method explored here can improve surface climate predictions over Europe.

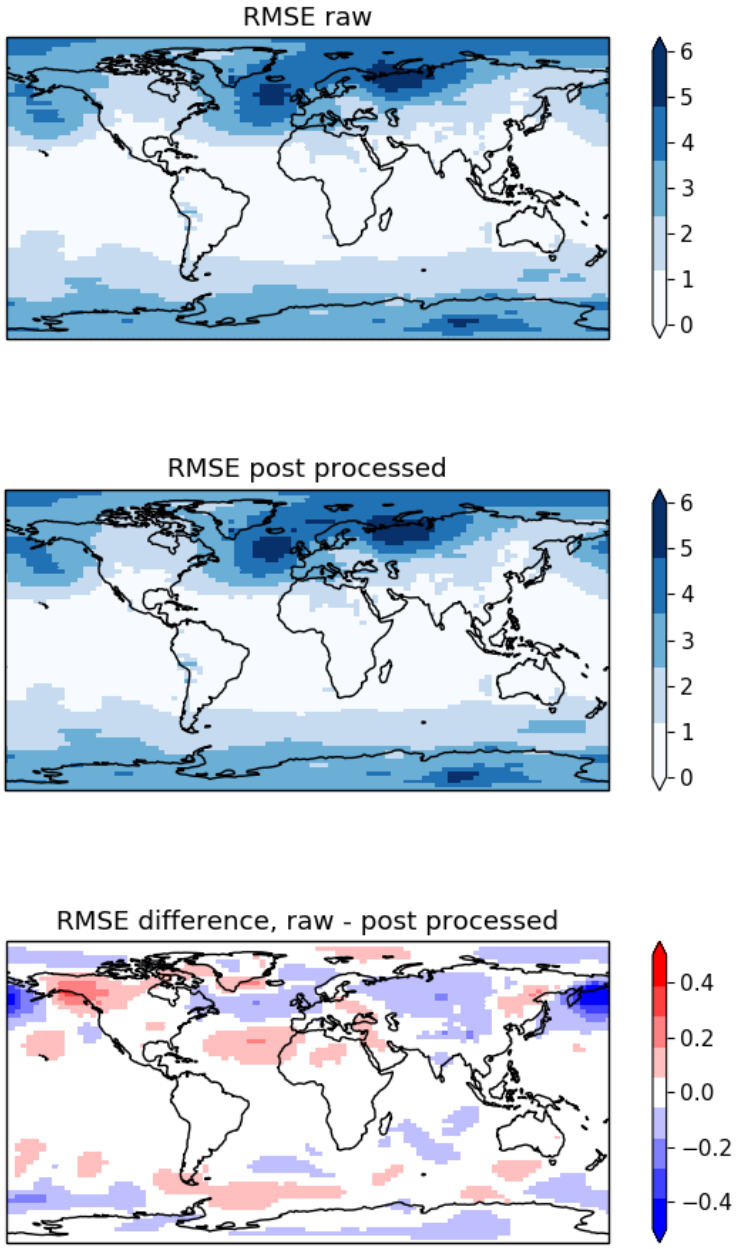

Figure 2.5.7: The Root Mean Square Error (RMSE) of MSLP for the raw multi-model ensemble mean (upper), the post-processed ensemble mean (middle) and the difference (lower). In the lower panel red shows the signal inflation/sub-selection method gives an improvement in forecast MSLP, blue a degradation.

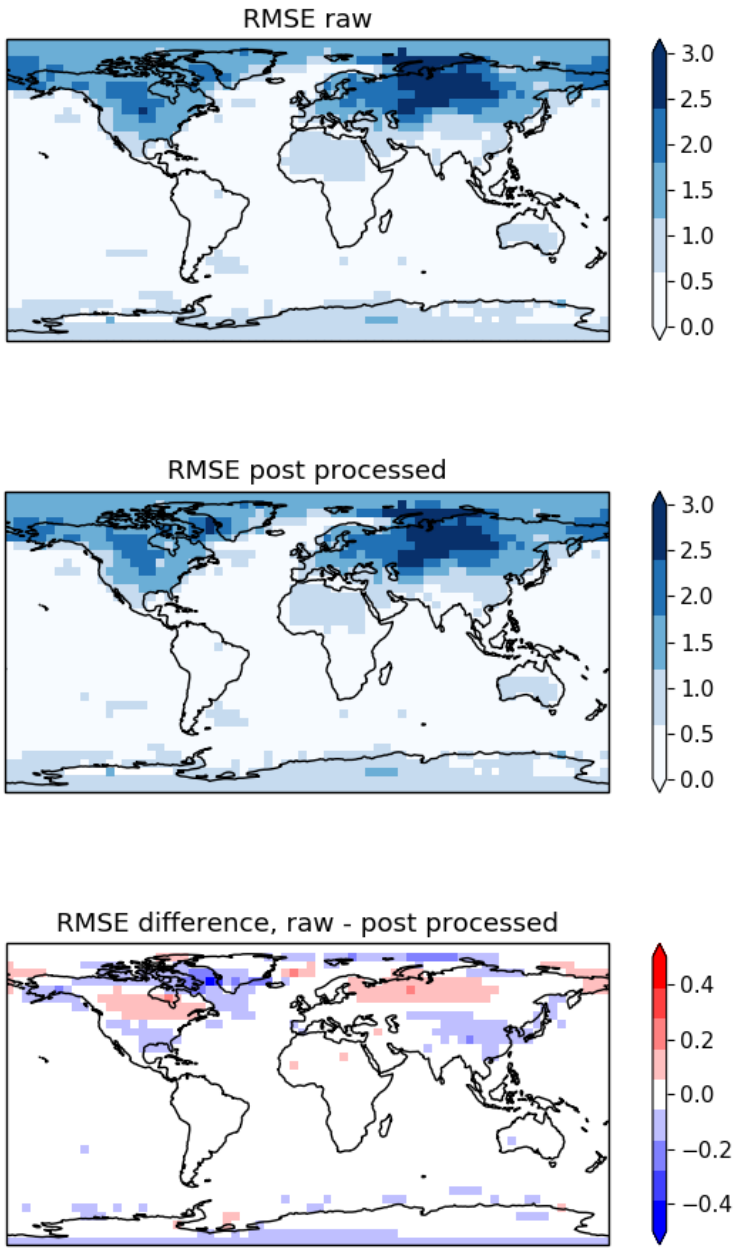

Figure 2.5.8: The Root Mean Square Error (RMSE) of 2m temperature for the raw multi-model ensemble mean (upper), the post-processed ensemble mean (middle) and the difference (lower). In the lower panel red shows the signal inflation/sub-selection method gives an improvement in forecast temperature, blue a degradation.

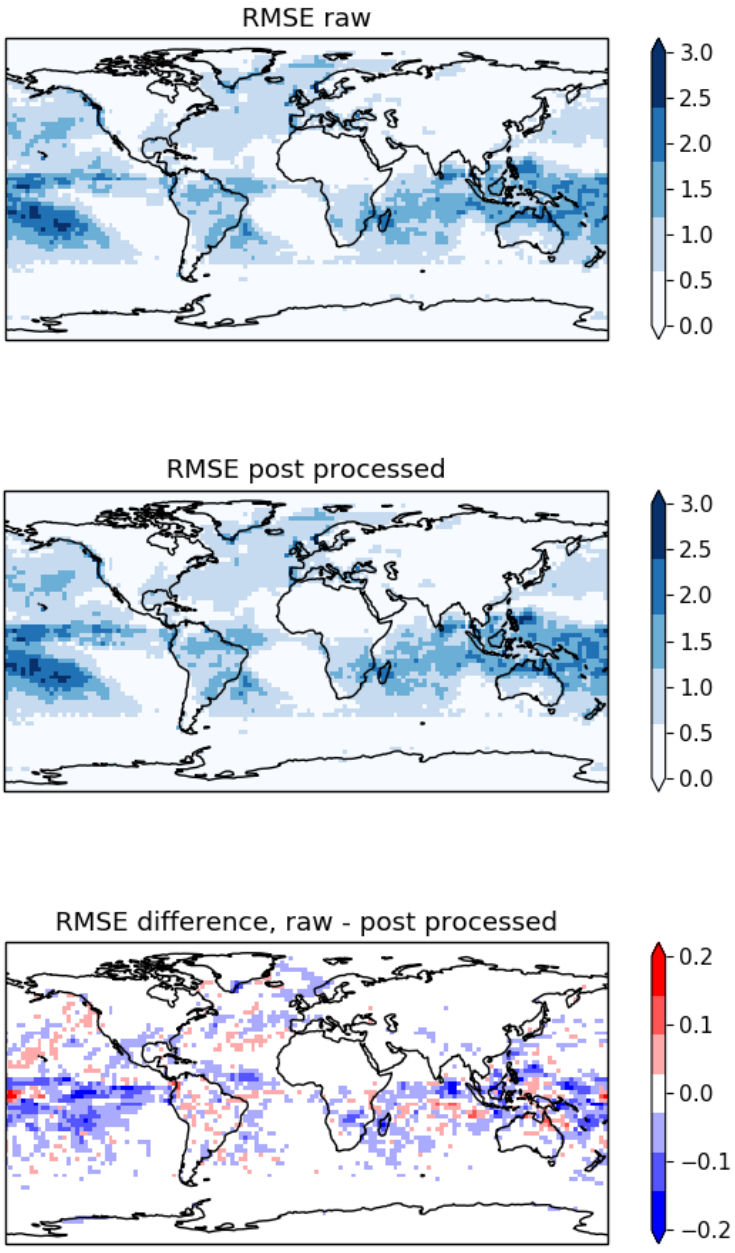

Figure 2.5.9: The Root Mean Square Error (RMSE) of winter mean precipitation rate for the raw multi-model ensemble mean (upper), the post-processed ensemble mean (middle) and the difference (lower). In the lower panel red shows the signal inflation/sub-selection method gives an improvement in forecast precipitation, blue a degradation.

2.6 Calibration Boost method

Seasonal Forecasts (SFs) systems provide a large set of forecasts for the same month, over the same geographic area. As an example, the ECMWF System 5 (Johnson et al. 2019) provides 51 realizations for each predicted date. From the modeller's point of view, this multi-choice product allows for the most objective picture of the possible evolutions of the weather in the future. Instead, from an end-user perspective, the requirement is binary: yes or no, whether to expect an adverse event or not. Industrial and financial decisions require non-ambiguous answers on whether to hedge or whether to insure against the anomaly.

To meet user requirements, a surrogate of probabilistic approach should be applied to the SF ensemble members to reduce multiple "opinions" between the ensemble members into a single value, or signal. Typically the standard choice in this case is to consider the ensemble mean. Here however we want to try to improve on this standard approach, as the ensemble mean tends to have a weak signal compared to the observed one. In this work we developed a methodology, called Calibration Boost (or Boosted Mean), to transform multiple SFs into a single and more pronounced signal. In essence, a sample of members is selected based on the confidence of the forecast. For example, if more than 60% of ensemble members agree on the sign of the anomaly, then these ensemble members are considered while those that present the opposite sign are rejected.

The methodology here presented, aims to obtain the method which represents best the ERA5 anomalies in terms of their month-to-month variability. The assessment of the Calibration Boost methodology is focused on various spatial scales: at local scale, globally for each model grid, "global area-weighted index" and, also at a country level. The agreement between calibrated SF and ERA5 is then evaluated at these spatial scales for each calendar month and each forecast lead time.

2.6.1 Data

For this investigation, we used SF outputs from five models: ECMWF, DWD, Météo-France (MF), NCEP and CMCC. We considered only monthly averages of three weather parameters: 2m air temperature (T2M), 10m wind speed (WS10) and the global downwelling short-wave horizontal irradiance (GHI). The ERA5 reanalysis (Copernicus Climate Change Service, 2017) was chosen as the reference for both the SF calibration and for the SF assessment. Both SF and ERA5 data are processed over the 24-year period, 1993-2016, namely the common hindcast period for SFs, and with a 1^{o} resolution. We consider 25 ensemble members for each model.

2.6.2 Methodology

The Calibration Boost methodology was developed through six variations. These depend on the method anomalies are computed. They are described next.

1. *SF_orig_anoms* refers to SF anomalies calculated as follows. The original 25 ensemble members are combined and the 23-year ERA5 climatology (namely, 24 years minus the one targeted) subtracted. To combine the 25 ensemble members we use the median of 25 values. 23-year ERA5 climatology is the multi-year median. When calculating SF anomaly, the corresponding multi-year climatology (reference for a given year) includes all years except the one being considered.
2. *SF_anoms*: to calculate SF anomalies we first apply a quantile correction to each of 25 ensemble members. Bias correction is applied to 25 SF ensemble members with respect to the ERA5 monthly distribution. Second: we take the median of these 25 corrected values. Third, we calculate the anomaly: SF median for current year minus 23-year median for ERA5. The reference climatology includes all years except the one being considered.
3. *SF_Boosted_anoms*: to calculate SF anomalies we first apply the quantile correction to each of 25 ensemble members. Second: we evaluated how many ensemble members suggest positive (negative) anomaly relative to reference SF long-term average (average out of 25 scenarios times 24 years). Next, if more than x% of ensemble members agree on the sign of anomaly, then we take the median of these "similar" forecasts, while rejecting the other SF ensemble members (rejecting (100-x)% of data). Thus, we give more weight to those forecasts with similar opinion. Last, we calculate the anomaly: SF for current year minus 23-year ERA5 climatology. This majority vote was tested with 60%, 70% and 80% thresholds. The reference climatology includes all years except the one being considered.
4. *SF_Boosted_anom_nqa_60*: to calculate SF anomalies we don't apply the quantile correction to 25 ensemble members. We apply the Boosting approach without a calibration. If more than 60% of ensemble members agree on the sign of the anomaly, then we take the median of these "similar" forecasts, while rejecting other SF ensemble members (rejecting less than 40% of data). 60% is the threshold for majority vote. Last, we calculate the anomaly: SF for current year minus 23-year SF climatology calculated from the original SF data. The reference climatology for each given year never includes current year, but all other years
5. *SF_Boosted_anom_nqa_70*: to calculate SF anomalies we don't apply quantile correction to 25 ensemble members. We apply "boosting" without prior calibration. Threshold is 70% majority vote. Last, we calculate the anomaly: SF average for current year minus 23-year median of the original SF data.
6. *SF_Boosted_anom_nqa_80*: to calculate SF anomalies we don't apply quantile correction to 25 ensemble members. We apply "boosting" without prior calibration. Threshold is 80% majority vote. Last, we calculate the anomaly: SF average for current year minus 23-year median of the original SF data. The reference climatology for each given year never includes current year, but all other years.

2.6.3 Results

We evaluated the agreement between each version of SF anomaly product (as listed above) and ERA5 at each grid location, for each calendar month, and lead times from 1 to 5 months.

All six methods were applied to the five SF models. We assess the skill of the seasonal forecasts by calculating the Correlation Coefficient, Adjusted R-Squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE). While we also processed WS10 and GHI, for brevity we show only the results for T2M.

2.6.3.1 Local comparison for 2m air temperature (TA)

We illustrate first the Calibration Boost methodology for one randomly selected grid location over land: 45°N 1°E (Figure 2.6.1). The distribution of all original 25 ensemble members in all years (for a given calendar month) in the context of ERA5 temperature distribution is shown in Figure 2.6.2.A. It illustrates the original SF data before applying the Calibration Boost method, and compares the distributions between ERA5 and the original SF data. At each given location we have only 24 monthly values of temperature in January for ERA5, while for ECMWF SF we have 24 years with 25 ensemble members for each calendar month. Thus, Figure 2.6.2.A compares the distribution of the 24 values for ERA5 versus the distribution of 600 (24x25) values for SF. The distribution is calculated as percentage of the available data, *i.e.* percentage out of 24 values for ERA5 or out of 600 values for SF. We observe some differences in the distributions, with narrower distribution of ERA5 against the SF data and a higher upper tail. Additionally, Figure 2.6.2.A indicates also that a 24-year record is too short for distribution corrections, as the 24-value sample of ERA5 is most likely not gaussian.

Figure 2.6.1. Focus domain for methodology test. Model (45°N 1°E) and the corresponding 1°lat by 1°lon grid box.

In Figure 2.6.2 B, C and D we compare ERA5 year-to-year monthly anomalies in January versus the 24 SF anomalies calculated with SF_orig_anoms, SF_anom (median of quantile adjusted ensemble members) and SF_B_anoms (quantile adjusted SF anomalies with the 70% majority vote). For B and C, we perform a quantile adjustment to correct (shift) each SF

ensemble member value according to the quantile range it falls into. In fact, we can observe that the calibration reduces the extreme SF ensemble members values reduces towards “less extreme” values, as the distribution of the resulting quantile adjusted SF monthly anomalies has no tails. However, when the majority vote is applied to the quantile adjusted SF ensemble members (Figure 2.6.2.D) it generates new positive anomalies. This method yields a ‘conservative’ distribution, namely with evenly distributed weak positive and negative anomalies. While such distribution of T2M does not follow a normal distribution, this result could still add a value for further classifications between near-normal and rare events. This seems to point to the fact that in order to achieve a more effective quantile adjustment a longer reference data set would be needed, perhaps using the entire ERA5 period (though this would introduce inconsistencies between ERA5 and SF, such as different trends, due to the different periods used).

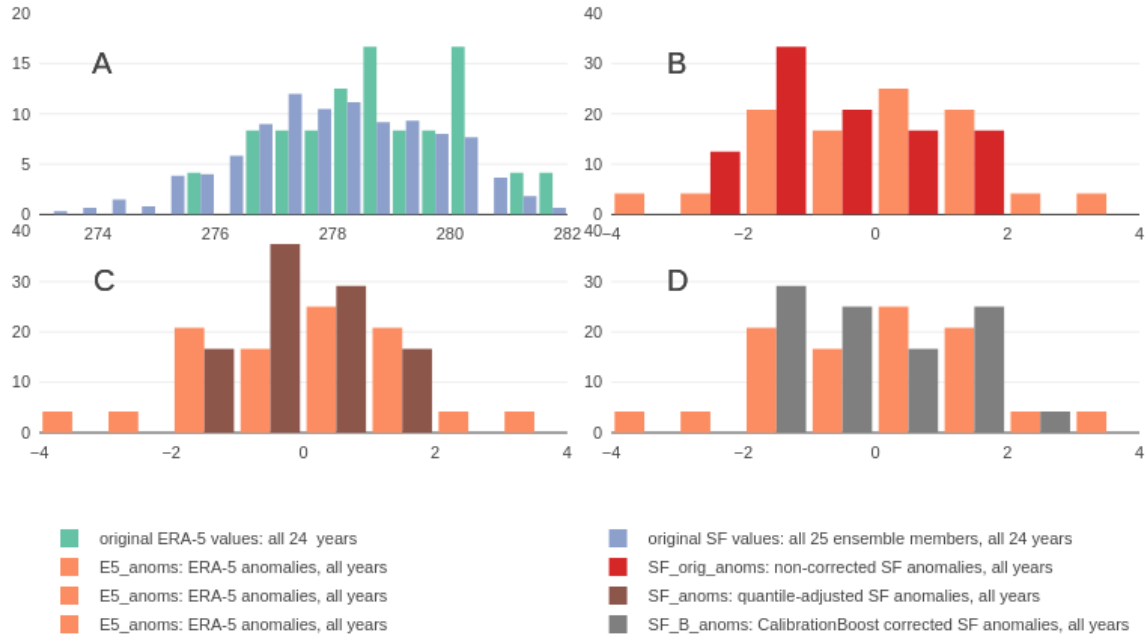

Figure 2.6.2. Comparison between ERA5 and SF: before and after calibration method applied. Starting month January, lead time zero month ahead, Forecast for January. Temperature forecast for January is compared against the historical record. (A) Distribution of the original ERA5 monthly temperature values (24 years) and the original ECMWF SF monthly temperature values with all ensemble members for each of 24 years. Units on plot A : [°K] Distributions of anomalies: (B) E5_anoms vs SF_orig_anoms, (C) E5_anoms vs SF_anoms, (D) E5_anoms vs SF_B_anoms. Anomalies [°K] are calculated relative to ERA5 climatology, location : 45°N 1°E.

While Figure 2.6.2 B-C-D reflects the distribution of temperature anomaly values during a 24-year period, Figure 2.6.3 illustrates the year-by-year evolution of the same temperature anomalies. Here we show the SF anomalies of the ECMWF model calculated with all 6 methods tested in this work. This example in one location demonstrates that only boosted methods (with majority vote) without quantile calibration (dashed curves) capture better the

amplitude of big anomalies of ERA5. This result shows, on the one hand, that if the user's objective is to capture anomaly amplitudes, the boosted method (with 60, 70, 80% majority vote) is very useful. On the other hand, it remarks the limitation of using a short sample data (24 year), as the quantile correction tends to underestimate the amplitude of year-to-year anomalies.

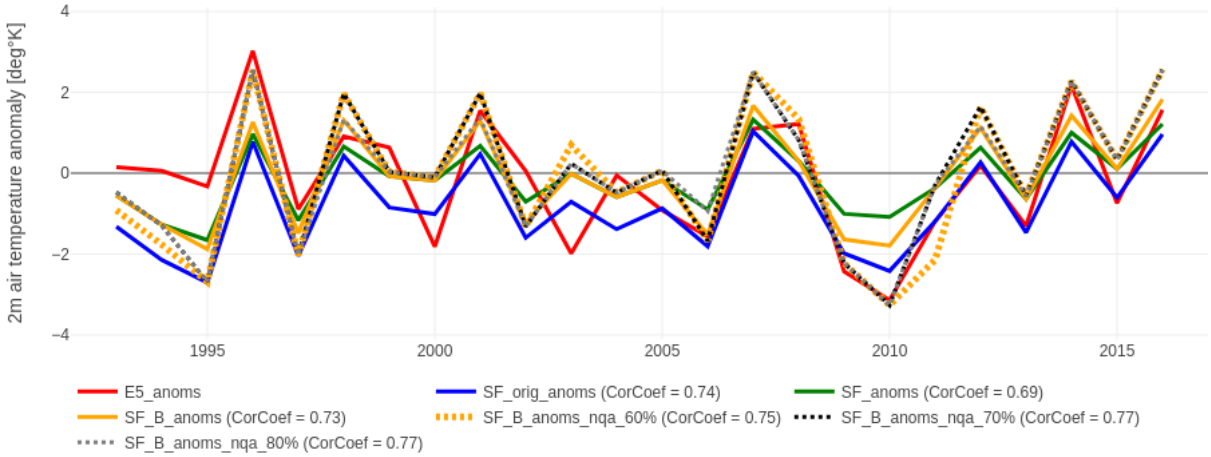

Figure 2.6.3. ECMWF SF and ERA5 2m air temperature year-to-year monthly anomalies. Starting month January, lead time 1 month ahead. SF anomalies are calculated with 6 methods. <nqa> abbreviation holds for <no quantile adjustment>. Dashed curves: <nqa> SF anomalies are calculated relative to SF climatology (except same year). Plain curves: SF anomalies calculated relative to ERA5 climatology (except same year). Both ECMWF SF and ERA5 are in 1°lat by 1°lon resolution. Location 45°N 1°E.

Success Score metrics were also applied to each of the six SF anomaly calculation methods versions (Figure 2.6.4). The scores were classified into four groups:

- Successful alert - when the forecasted alert is correct in both, sign and amplitude. It is measured as percentage of such alerts relative to all alerts which should have been done.
- False alert - when the alert is forecasted by SF, while nothing happened. It is measured as percentage of such alerts relative to all alerts ever done, no matter the anomaly sign (because the anomaly didn't happen).
- Missed alert - when the extreme happened, extreme not predicted by SF. It could happen that forecasted anomaly is too weak (i.e. prediction of non-extreme event). It is measured as percentage of such alerts relative to all alerts which should have been done.
- Wrong alert - when the extreme happened, extreme predicted BUT predicted with a wrong sign. It is measured as percentage of such events relative to all alerts ever done.

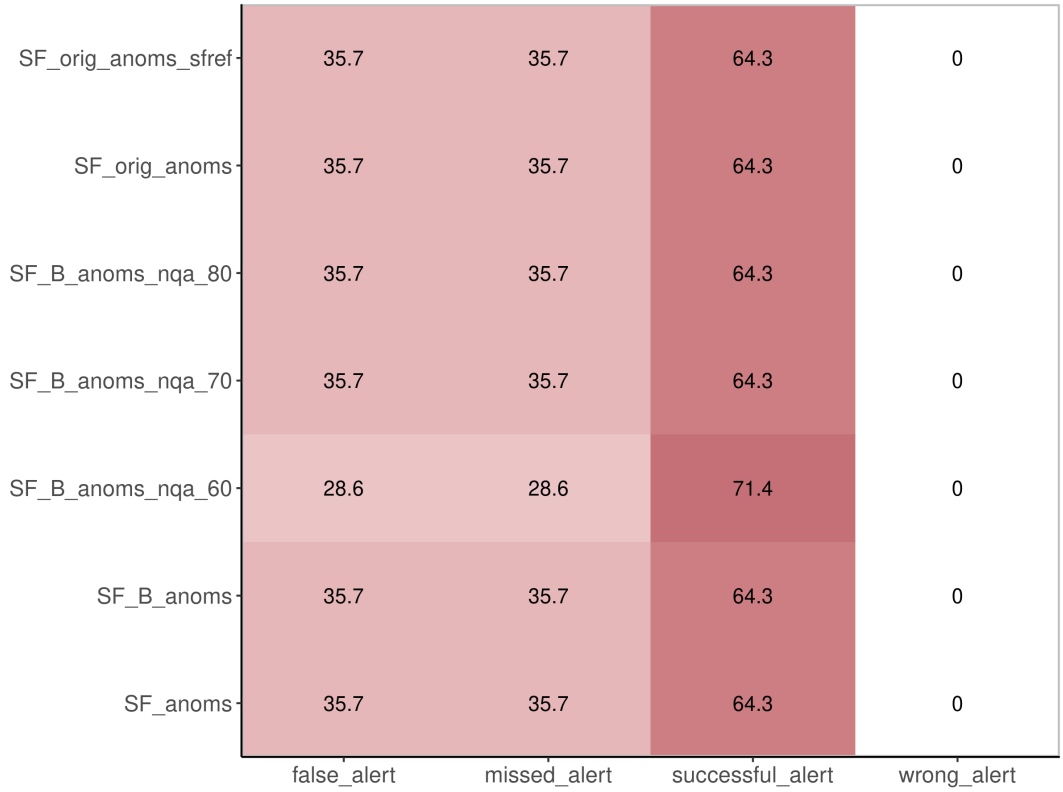

Figure 2.6.4. Success score (%) for anomaly forecast for both lower and upper 30% distribution tails together. Best results are highlighted in colour. SF data: ECMWF. Starting month January. Leadtime 1 month. Having a 24-year record (1993-2016), the upper and lower 30% tails represent together 14 years. <nqa> abbreviation stands for <no quantile adjustment>. <nqa> SF anomalies are calculated relative to SF climatology (except same year). <SF_orig_anoms_sfref> anomalies are calculated relative to SF climatology also. Anomalies for other SF versions are calculated relative to ERA5 climatology (except same year). Location: 45N 1E.

At this particular location, about 64-71% of extreme events were forecasted to be extreme by all six methods, with the best performance of the "60% majority vote" without quantile adjustment (SF_Boosted_anom_nqa_60). Also, at this location there were no wrong alerts (the extreme event happened according to ERA5 and it was predicted with the wrong sign). False alerts occurred between 28-36% of SF alerts. This could indicate to the possibility of having an anomaly in ERA5 that year, but not an exceptional anomaly predicted by the SF. Finally, 29-36% of extreme anomalies were registered as missed alerts, SF not forecasting as extremes compared to ERA5 by either method.

We want to remark that the definition of the threshold for extreme events depends on the definition with regard to the final user. For instance, it could be defined in respect to some economic indicator or yield volume loss. In this study we used the arbitrary anomaly threshold to illustrate the skill of the Seasonal Forecasts over the test area.

2.6.3.2 *Analysis of results: skill of Seasonal Forecasts for 2m air temperature (TA) on a global scale.*

On a global scale and for all calendar months, SF_orig_anom method correlates best with ERA5 compared to the other five SF anomaly calculation methods. An example of the regional differences in the correlation coefficient for the forecast of February is shown in Figure 2.6.5. In general, we can conclude that in those regions where the correlation coefficient is above 0.5 there is an added value of the seasonal forecasts produced in January for one month ahead. For T2M SF_orig_anom method is the only one giving significant correlation for lead time 2, regardless of the SF model.

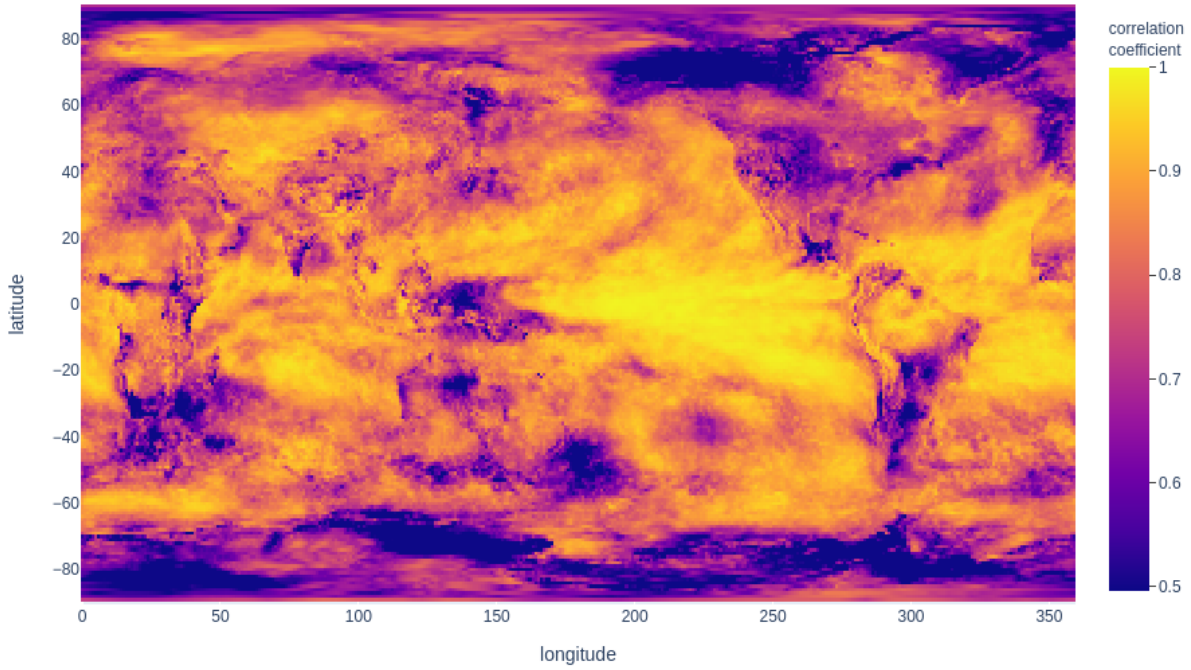

Figure 2.6.5. Comparison between ERA5 and SF_orig_anom. Correlation Coefficient is calculated at each grid location. SF data: ECMWF. Starting date is January with lead time 1 month (forecast for February initialised in January). Correlation Coefficient calculated with a 24-length sample is significant at 99% if exceeds 0.496. Correlations are significant in all areas except those highlighted by dark blue colour.

Figure 2.6.6 illustrates the month-by-month evolution of the globally averaged skill of the SFs trough the Adjusted R-Squared at each grid point and averaged globally. For each starting forecast month (calendar month) we have six forward forecast steps (lead times), corresponding to the six months ahead. The first forecast (six-month long curve) is the Adjusted R-Squared for each of six months between January and June. The second 6-month long forecast is the Adjusted R-Squared for months from February to July. And so on until the last forecast, corresponding to months from December to May of the next year. As an example, ECMWF SF_orig_anom explains up to 37% of variance 2 months ahead (lead time 2), while

other five SF anomaly calculation methods explain up to 32% of variance in ERA5 monthly temperature anomalies.

On a global scale the agreement (Adjusted R-Squared) between SF and ERA5 is high for the first forecast month (Figure 2.6.6), dropping by as much as 50% for the following forecast month (lead time 2). Among the six SF calculation methods, SF_orig_anom performs best throughout the year. The spread between the six points (for the same calendar months) demonstrates the sensitivity of the results to the method chosen. According to this global picture: while the method used to calculate SF_orig_anom allows for a three and even four months forecast for all calendar months, other methods do not perform that well. This conclusion doesn't necessarily hold on a local scale.

For lead time 3 on a global scale, ECMWF SF_orig_anom still performs well, explaining along the entire year between 26-32% of variance in ERA5 monthly temperature anomalies (Figure 2.6.6). These encouraging results on a global scale indicate that ECMWF SF_orig_anom could be a good option for forecasting 3 months ahead in all calendar months, all seasons, both on a local and regional scales. When the DWD system is tested, its results agree with ECMWF on global scale. There is also potential in testing DWD SF_orig_anom on a regional scale up to 3 months ahead. On the global scale, with the exception for April, May and October, DWD SF_orig_anom with the lead time of 3 months ahead explains 24-27% of variance (significant at 99%) of ERA5 monthly temperature anomalies. On a global scale our results for the Météo-France model indicate that SF_orig_anom could be used for a 3 month ahead forecast in all months except for the starting months from March-to-June. On a global scale the skill of NCEP model for up to 3 months ahead is weaker compared to other models explored here, independent of the calendar month.

On a global scale, the forecasts by ECMWF and CMCC models appear as the most relevant for 1 to 3 months ahead. CMCC SF_orig_anom explains up to 54%, 32% and 28% of variance respectively for 1 to 3 months ahead (significant at 99% level). These results obtained from a global scale on average, and may differ from one geographic zone to another.

In terms of the Mean Absolute Error (MAE) of seasonal forecasts, when comparing SF anomalies and ERA5 on a global scale, the ECMWF model is of the order of magnitude 0.5-1.2°C between different calendar months and different lead times. On a global scale, for all models, MAE is smaller during May-October months.

2.6.4 Evaluation of Calibration Boost methodology: results case study 4

After the local and the global assessment of results we focus on regional performance of Calibration Boost methodology. Seasonal forecasts are evaluated at each model grid location within the continental Spain. Figure 2.6.7 illustrates the map of the Adjusted R-Squared at each model grid location (numbers). Year-to-year monthly SF anomalies for each model grid location are calculated relative to local ERA5 climatology. As the figure shows, the highest variability explained is located in the west of Spain, and particularly north-west, driven by large fronts coming from the Atlantic that might be easier to predict.

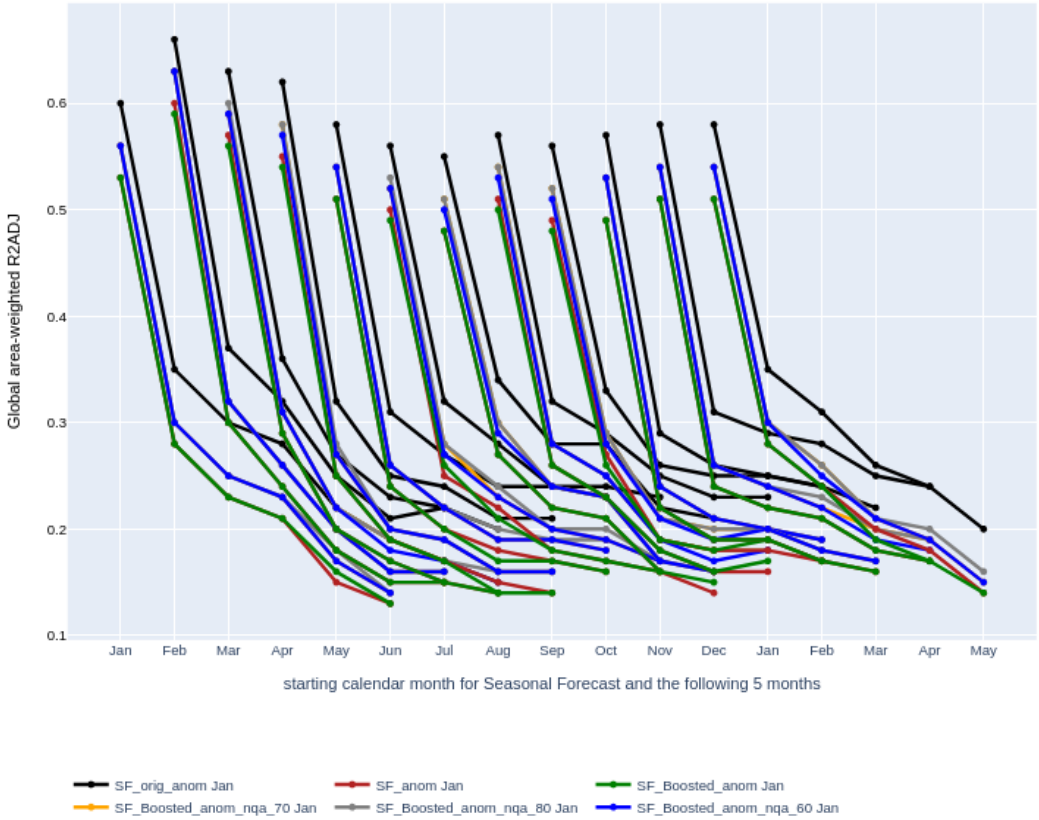

Figure 2.6.6. Month-by-month evolution of the globally averaged Adjusted R-squared calculated for each model grid. Adjusted R-squared statistics is calculated between ERA5 monthly anomalies and 6 versions of SF anomalies. Parameter: 2m air temperature. SF data: ECMWF. Each local value of Adjusted R-squared is first grid box area weighted, and then after area averaged over the entire Globe without any filters. Grid resolution 1°lat by 1°lon. For each given starting month: six curves correspond to either version of SF anomaly calculation methods. <nqa> abbreviation holds for <no quantile adjustment>. No filter applied in order to preserve the global picture of the SF skill. No significance test applied. No normal distribution test applied. Years: 1993-2016

Figure 2.6.8 illustrates the adjusted R-Squared as the proportion of the year-to-year variance in ERA5 temperature anomalies explained by year-to-year variations in SF monthly temperature anomalies. It is shown how results depend on the season (calendar month), the forward forecast time scale (up to six months ahead) and SF anomaly calculation methodology (six versions of SF anomaly).

On the left-hand side of Figure 2.6.8 we have the December forecast for 6 months ahead (for January-June months). This December forecast is illustrated with six curves (one for each method) all correspond to the forecast starting month in January. Those forecasts with the starting month in January are referenced in the legend as “Jan”. Forecasts with the starting month in December (valid for December-May next year) are referenced in the legend as “Dec”. For each starting calendar month and each forecast lead time step (1 to 6 months), we

calculate the regional average of all local Adjusted R-Squared within country contour. The average of all local Adjusted R-Squared values showed in Figure 2.6.7 corresponds to one value of the “SF_orig_anom” curve in Figure 2.6.8. The regional average Adjusted R-Squared for August as starting month and one month lead time is equal to 0.45. The regional average in this example is evaluated within the continental Spain.

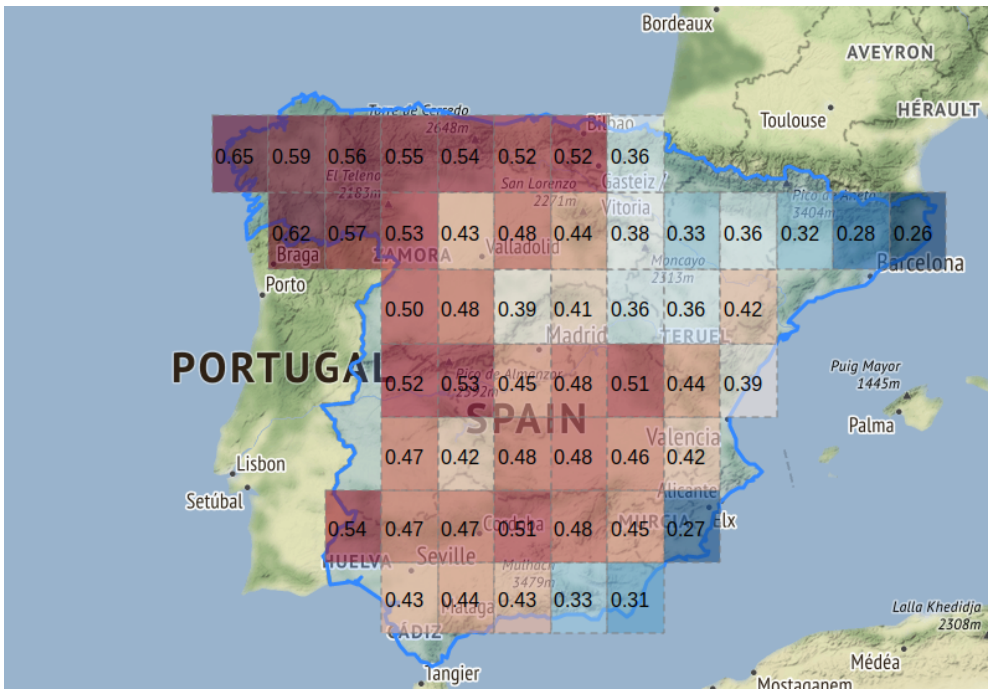

Figure 2.6.7. Adjusted R-Squared (numbers) between ERA5 and ECMWF SF_orig_anoms anomalies for 2m air temperature. Starting month: August. Leadtime: 1 month ahead. Adjusted R-Squared (agreement between SF vs ERA5) is calculated individually for each 1°lat x 1°lon grid location. The median is used to calculate the reference climatology and also for the ensemble member averaging. In this example, the overall average of local Adjusted R-Squared values is equal to 0.45. This result indicates that, on average, the Seasonal Forecast (in August) for September explains 45% of the year-to-year variations in ERA5 anomalies in September, while locally the Adjusted R-Squared values range within 0.26-0.65. For a 24-year record the Adjusted R-Squared is significant if exceeds 0.24.

Results differ a lot depending on the method used, the selected region, the starting calendar month, and the forward forecast time step. As can be deduced from Figure 2.6.8, there is a spread between the six anomaly calculation methods. SF_orig_anom is calculated with the simplest method and also it appears to have the highest Adjusted R-Squared values when comparing to ERA5: highest relative to other five anomaly calculation methods. When comparing between five SF models, the explained variance for 1-month lead time is significant and the highest for ECMWF model. This statement holds for this particular region, while the model performance could be different between regions and different weather parameters. According to our results, the explained variance for ECMWF for 1-month lead time is within 28-62% though the entire year, above 50% for January, April, May, July and September months.

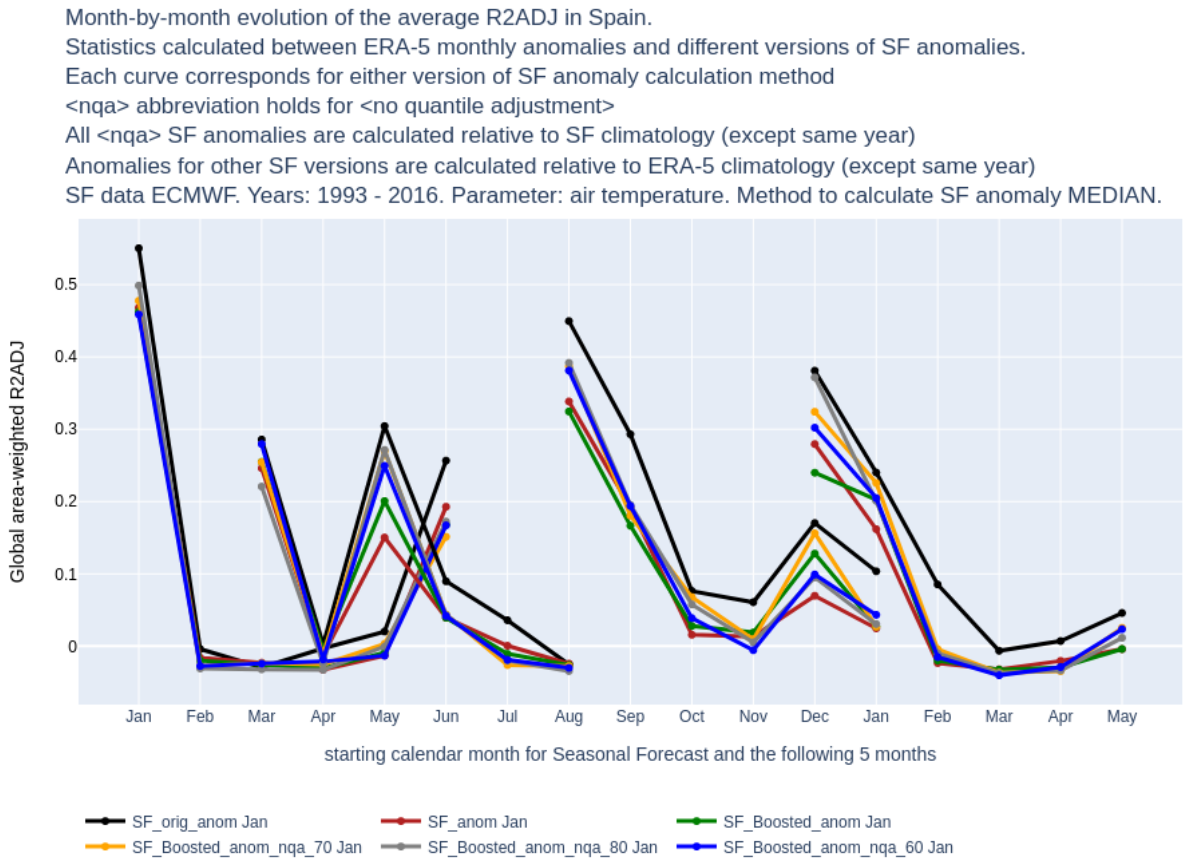

Figure 2.6.8. Adjusted R-Squared: comparison between ERA5 and ECMWF SF anomalies. Parameter: T2M. Example for starting months (first forecast month): January, March, August and December. Forecast lead times are one to six months ahead which explains why the lines are six-months long. Adjusted R-Squared values are calculated individually for each 1°lat x 1°lon grid point and then averaged over the region. For a 24-year record the Adjusted R-Squared is significant if exceeds 0.24.

Comparison for this region between different SF models can be summarised as follows. The explained variance for the CMCC model for one month lead time is within 25-49%, so below 50% in all calendar months (weaker skill compared to ECMWF model). Explained variance for DWD model for 1-month lead time is within 0.18-0.52, only above 50% for starting month January. Explained variance for Météo-France model for 1-month lead time is within 0.11-0.49 in all months (Figure 2.6.9). Explained variance for NCEP model for 1-month lead time is within 0.10-0.37 in all months.

In Figure 2.6.10 is shown the evolution of the Mean Absolute Error of SF_orig_anom throughout the year of the Météo-France model, which have similar values for the different initialized months. It is also remarkable from this plot, that depending on the initialization month, the lead time dependency on the MAE it is not very strong. Instead, there seems to be more related to the predicted month itself, showing larger values in the winter months than in the spring-summer months.

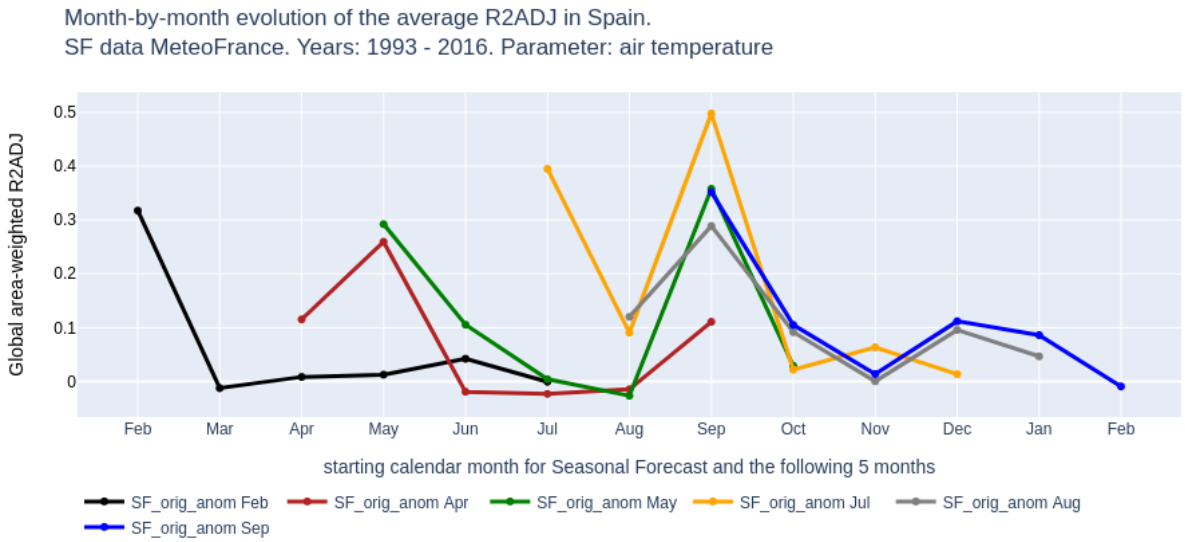

Figure 2.6.9. Adjusted R-Squared between ERA5 and Météo-France SF anomalies. Parameter: 2m air temperature. Example for starting months (first forecast month): February, April, May, July, August and September. Forecast lead times are one to six months ahead. Adjusted R-Squared are calculated individually for each 1°lat x 1°lon grid point. For a 24-year record the Adjusted R-Squared is significant if exceeds 0.24.

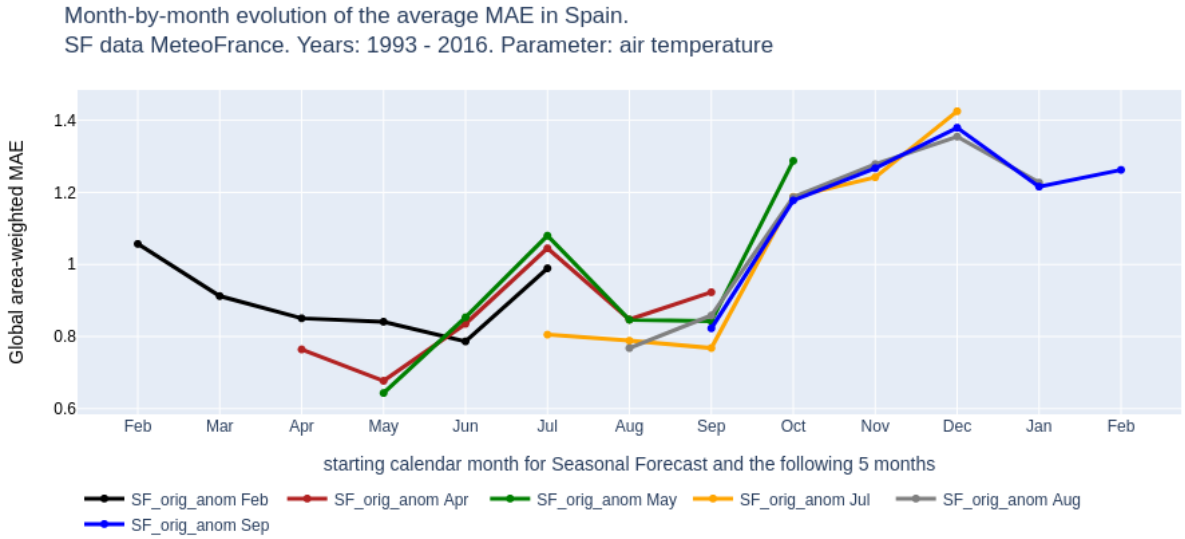

Figure 2.6.10. Mean Absolute Error (MAE) for Météo-France model: comparison of SF anomalies relative to ERA5 anomalies. Parameter: 2m air temperature. Units: °C. Example for starting months (first forecast month): February, April, May, July, August and September. Forecast lead times are one to six months ahead: 6-month long curves. Mean Absolute Error (MAE) is calculated individually for each 1°lat x 1°lon grid point and then averaged within the region.

Results for five models agree that the skill does not decrease linearly with the forward forecast time step (Figure 2.6.8 and Figure 2.6.9). Following ECMWF model (Figure 2.6.8) the December forecast is significant for January and also for June with an Adjusted R-Squared above 0.1, while months in between obtained values close to zero (at least with the current method formulation, given area). Also, according to Figure 2.6.8 the ECMWF February forecast works well for March and May, while non-relevant for April. Météo-France model (Figure 2.6.9) follows a similar pattern. Météo-France forecast in June (starting month is July) is significant for July and September, while non-relevant for August (Figure 2.6.9).

Figure 2.6.11 illustrates the Adjusted R-Squared for the same starting month but for the forward forecast of 2 months ahead instead of 1 month as illustrated in Figure 2.6.7.

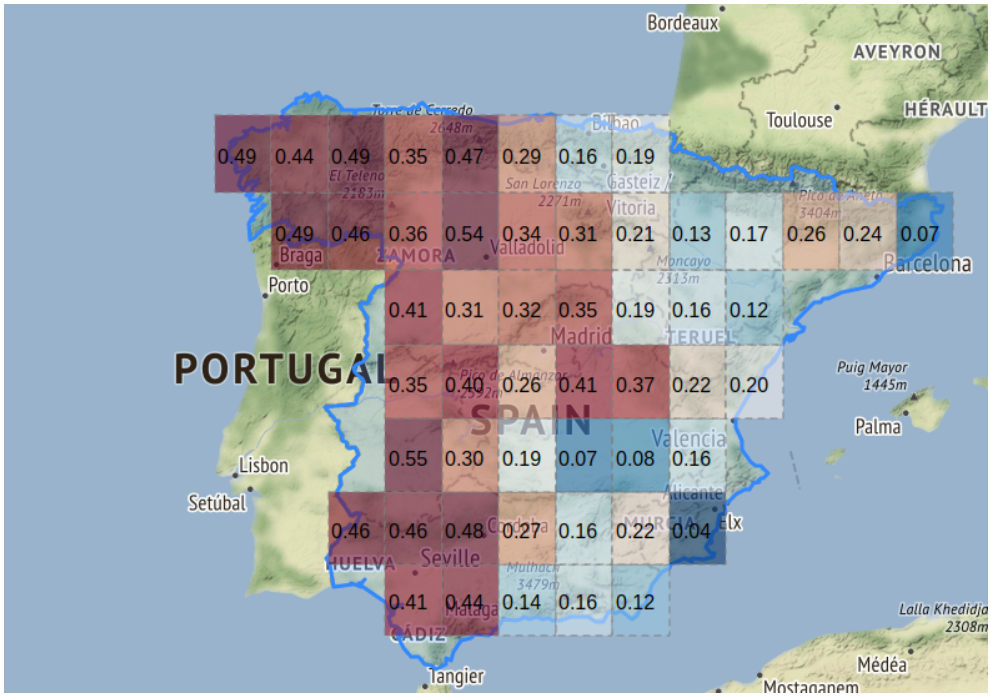

Figure 2.6.11. Adjusted R-Squared (agreement) between ERA5 and ECMWF SF anomalies. Parameter: 2m air temperature. Adj. R-Squared is calculated individually for each 1°lat x 1°lon grid point. Contour: continental Spain. Starting month: August. Leadtime: 2 months ahead. Method: median for climatology calculations and also for ensemble member averaging. In this example, the overall average of all local Adj. R-Squared values is equal to 0.29. This result indicates that, on average, the Seasonal Forecast (in August) for September explains 29% of the year-to-year variations in ERA5 anomalies in September, while locally the Adj. R-Squared values range within 0.04-0.55. For a 24-year record the Adj. R-Squared is significant if exceeds 0.24.

2.6.5 Conclusions

1. Quantile adjustment alone has a positive effect on bias correction. So far quantile adjustment generates a loss in variability between the original SF anomalies and the corrected SF anomalies. True for low resolution ERA5 in 1°lat x 1°lon grid.
2. Seasonal Forecast anomalies calculated with the Calibration Boost method ("majority vote") applied to filter between ensemble members appeared to be useful. SF anomalies calculated with the "majority vote" capture well the anomaly amplitudes and variability.
3. Shifting the 23-year sample for climatology calculations contributes to uncertainty in year-to-year anomaly amplitude within 0.3°C. Year-to-year anomaly calculation relative to 23-year climatology depends on the year choice for climatology. In order to reduce the effect of year sampling, 23-year climatology should be calculated as a median of 23 values instead of the mean of 23 values. Median method provides more conservative (more stable) climatology when including/excluding different years from a 23-year sample.
4. A 23-year-long ERA5 sample is enough for overall bias adjustment in SF data but is a limitation for representation of extreme events. Thus, we observe the loss of extremes when SF data are calibrated against a 23-yearlong sample of "local" monthly ERA5 climatology deduced from 1°lat x 1°lon grid. Two possible improvements: to use ERA5 in 0.25° spatial resolution instead on 1-degree resolution, or to use a longer ERA5 record. 23-yearlong sample is not enough representative for year-to-year anomalies.
5. In this work the goodness of results was measured with the Correlation Coefficient, Adjusted R-Squared, MAE, MSE, RMSE and the Success Score. Threshold for the "goodness" will depend on further user requirements.
6. For lead time 2 months ahead only SF_orig_anoms anomaly calculation method appears to be useful. To note, it is the simplest method out of the 6 tested here. It's important to note that the performance of quantile adjustment method should improve if tested for a longer data sample.
7. When comparing between five models (TA parameter), the explained variance for 1-month lead time is significant and the highest for ECMWF, followed by CMCC model.
8. On a global scale, ECMWF "SF_orig_anom" performs well for lead time 3. Along the entire year ECMWF SF_orig_anom with 3 months lead time explains 26-32% of variance in ERA5 monthly temperature anomalies.
9. When shaping a regional indicator, an appropriate combination of grid points could be selected using MAE thresholds (MSE, RMSE) to remove locations with weak performance of Seasonal Forecasts.
10. When testing the new methodology, the majority vote method could also consider the amplitude of the anomaly and not only the sign of the anomaly as it is now. In other words, when many ensemble members (for example, at least 30% of them) suggest that the amplitude of the anomaly, for example, is greater than one standard deviation, then these ensemble members could have a larger weight compared to "conservative" (near-average) ensemble members.

2.7 Direct application of the Calibration Boost method to the end-user

The Calibration Boost methodology is also referred to as “The boosted mean method” in the specific documents for each case study (see D3.2, D3.3, D3.4, D3.5, D3.6) and it is applied in place of the previous “weighted mean method”. With the previous method, the weather forecasts showed very small differences with respect to the historical mean. Consequently, the added value of seasonal forecasts on Enel’s decision-making process would be very small or negligible. For this reason, a preliminary version of the boosted mean solution has been applied on Enel’s case studies in order to boost the signal of extreme events detected by the forecast models.

For Case Studies 1, 3 and 5 this method is used to derive the multi model results, while the single model approach utilizes the weighed mean method. The preliminary procedure is outlined as follows: the boosted mean approach has been computed for each involved model of seasonal forecast (FCST), each variable, each target time, and each forecast starting month. The spatially aggregated seasonal forecasts and ERA5 historical data are used to compute the 10th, the 33rd, the 50th, the 66th and 90th climatological percentiles of all datasets. These are respectively, for seasonal forecasts and ERA5 (or IDEAM stations in case study 5): $P_{10,FCST}$, $P_{33,FCST}$, $P_{50,FCST}$, $P_{66,FCST}$ and $P_{90,FCST}$ and $P_{10,ERA5}$, $P_{33,ERA5}$, $P_{50,ERA5}$, $P_{66,ERA5}$ and $P_{90,ERA5}$. These percentiles are computed for each model and for all months of the reference period 1993-2014, using an empirical, cumulative probability distribution function i.e. built on an existing dataset, (Figure 2.7.1).

This part has already been described in more detail in D3.12 and here: https://it.mathworks.com/help/stats/quantiles-and-percentiles.html. Each percentile of seasonal forecasts is subtracted by the corresponding percentile of ERA5 in order to calculate the value of bias adjustment for each probability interval. For example, if the value of such ensemble is greater than the 90th percentile, the bias correction is computed with the higher value of bias (or lower in case of 10th percentiles). Similarly, if it falls between two percentiles (10th to 33rd, 33rd to 50th, 50th to 66th or 66th to 90th) the average of the two extremes is subtracted.

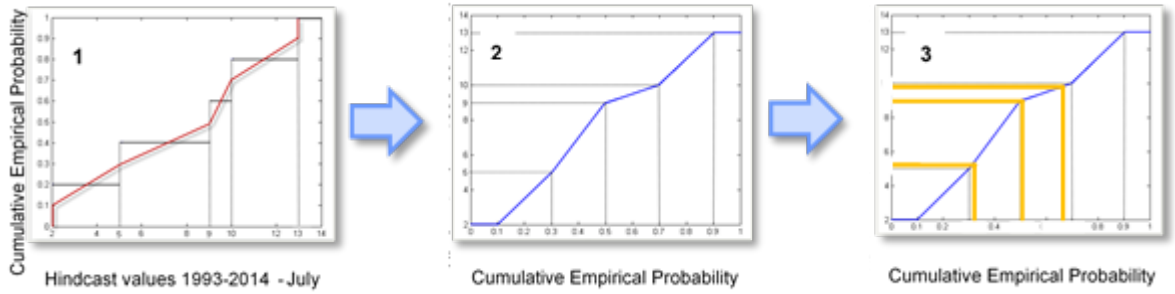

Figure 2.7.1: Example of process to derive the empirical cumulative distribution function derived by MATLAB software through the percentile function. Pictures modified from: https://it.mathworks.com/help/stats/quantiles-and-percentiles.html

Figure 2.7.2 shows an example of bias adjustment for temperature with a distribution of forecasts obtained from different ensemble members (red) and historical data from ERA5

(blue). This step allows us to scale climate model outputs to account for their systematic errors, in order to improve their fitting with the ERA5 model, which represents the best estimation of the real world.

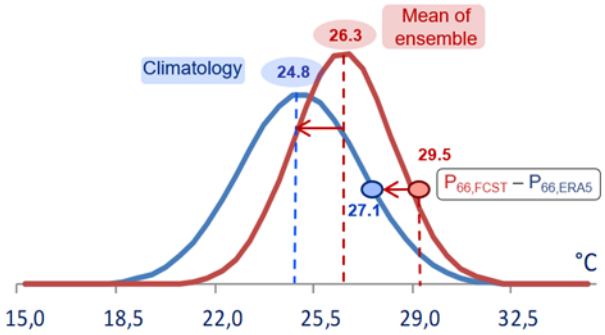

Figure 2.7.2: Example of bias adjustment for temperature. Seasonal forecasts (red), hindcast (blue)

The probabilities over seasonal forecast models are derived by counting the number of ensemble members for each model that fall below, above, or between percentiles and dividing them by the total number of ensemble members.

If 70% or more ensemble members fall below (or above) the 50th percentile ($P_{50,FCST}$), the algorithm detects an extreme event and the final forecast is computed with the median of the forecast ensemble members that fall below (or above) $P_{50,FCST}$. If the percentage of ensemble members below (or above) $P_{50,FCST}$ does not reach the 70%, no extreme events registered, and the final forecast corresponds to the median of all the forecast ensemble members**.** Figure 2.7.3 shows how the boosted mean algorithm works: on the picture on the left, no extreme event is registered; the amount of ensemble members that falls above/below $P_{50,FCST}$ does not reach the threshold value of 70%. While on the right, more of 70% of ensemble members fall above the $P_{50,FCST}$.In this case, an extreme hot weather event is detected.

The result of boosted mean method consists of four forecast values obtained from each model selected by WP2 (DWD, UKMO, MF, ECMWF). The value used in the Multi Model approach is the simple average of the four boosted mean values

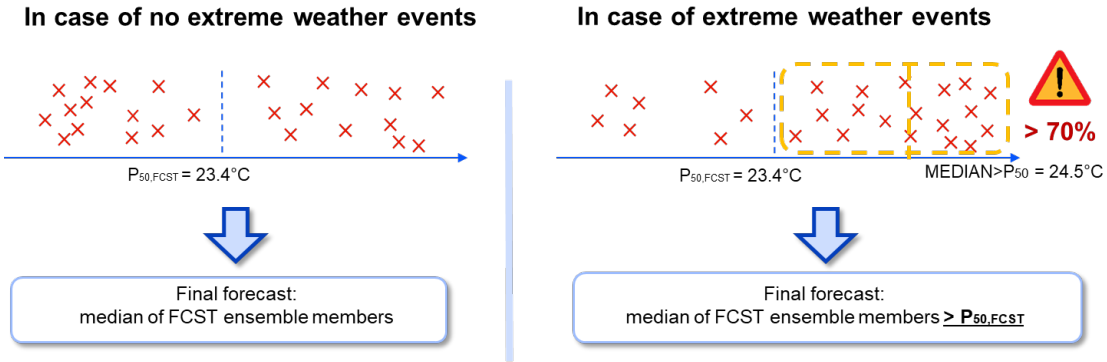

Figure 2.7.3: Example of the algorithm used to get the final forecast (boosted forecast) in case of no extreme weather events (left) and extreme weather events (right).

The boosted mean and the weighed mean are signal boosting methodologies useful for Enel to transform probabilistic information into a single deterministic value that feeds the market models. In particular, a forecast value as close as possible to the actual value leads to a better estimation of commodity exposures and a correct choice of the strategy to adopt.
In the Secli-firm case studies, the use of the boosted mean did not bring the expected results since the variation with respect to the weighted mean were too small to have an impact on the decision-making tree results.

Meanwhile, the WP2, has updated the boosted mean method as reported in section 2.6 to improve its performances for all variables.

2.8 Impact of North Atlantic Weather Regimes in the downscaling process

A quantile-mapping correction conditioned on the Weather Regimes (called ADAMONT) has been used in several configurations to study the impact of North Atlantic teleconnections on the quality of the downscaling. Although this technique relies on the role of large-scale climate phenomena on predictability, topic of this deliverable, the major strength of this approach is the use of weather regimes associated with deliverable D2.3 (Report on the predictability of weather patterns and regimes of relevance for the case study applications). Hence, the results obtained by Météo-France are presented in Deliverable 2.3.

3 References

Alessandri A., Catalano F., De Felice M., van den Hurk B., Doblas-Reyes F., Boussetta S., Balsamo G., Miller P. A., 2017: Multi-scale enhancement of climate prediction over land by increasing the model sensitivity to vegetation variability in EC-Earth. Clim. Dyn., 49, 1215-1237, doi:10.1007/s00382-016-3372-4

Alessandri, Andrea, Matteo De Felice, Franco Catalano, June-Yi Lee, Bin Wang, Doo Young Lee, Jin-Ho Yoo, and Antije Weisheimer. 2018. "Grand European and Asian-Pacific Multi-Model Seasonal Forecasts: Maximization of Skill and of Potential Economical Value to End-Users." Climate Dynamics 50 (7–8): 2719–38. https://doi.org/10.1007/s00382-017-3766-y.

Alessandri, A., Catalano F., Nielsen K., Troccoli A.: Grand multi-model seasonal forecasts for the energy industry. (in preparation).

Catalano F., Alessandri A., Nielsen K., Troccoli A.: A novel model independence methodology to improve multi-model seasonal forecasts combination. (in preparation).

Barnston, A. G., Glantz, M. H., and He, Y.: Predictive Skill of Statistical and Dynamical Climate Models in SST Forecasts during the 1997–98 El Niño Episode and the 1998 La Niña Onset, 80, 217–244, https://doi.org/10.1175/1520-0477(1999)080<0217:PSOSAD>2.0.CO;2, 1999.

Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global 355 climate. Copernicus Climate Change Service Climate Data Store (CDS), Date accessed: 15-01-2021.https://cds.climate.copernicus.eu/cdsapp#!/home

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models, 19, 275–291, https://doi.org/10.5194/hess-19-275-2015, 2015.

Eade, R. et al. Do seasonal to decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.* **41**, 5620–5628 (2014).

Eden, J. M., Oldenborgh, G. J. van, Hawkins, E., and Suckling, E. B.: A global empirical system for probabilistic seasonal climate prediction, 8, 3947–3973, https://doi.org/10.5194/gmd-8-3947-2015, 2015.

Fan, Y. and Dool, H. van den: A global monthly land surface air temperature analysis for 1948–present, 113, https://doi.org/10.1029/2007JD008470, 2008.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Silva, A. M. da, Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W.,

Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B.: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), 30, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1, 2017.

Hagedorn, Renate, Francisco J. Doblas-Reyes, and T. N. Palmer. 2005. “The Rationale behind the Success of Multi-Model Ensembles in Seasonal Forecasting - I. Basic Concept.” Tellus A 57 (3): 219–33. https://doi.org/10.1111/j.1600-0870.2005.00103.x.

Harris, I., Jones, P. D., Osborn, T. J., and Lister, D. H.: Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset, 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

Hersbach et al (2018) Operational global reanalysis: progress, future directions and synergies with NWP. ERA Report series, ECMWF, 65 pp.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons, 30, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1, 2017.

Johnson, S. J., Stockdale, T. N., Ferranti, L., Balmaseda, M. A., Molteni, F., Magnusson, L., Tietsche, S., Decremer, D., Weisheimer, A., Balsamo, G., Keeley, S. P. E., Mogensen, K., Zuo, H., and Monge-Sanz, B. M.: SEAS5, 2019: the new ECMWF seasonal forecast system, Geosci. Model Dev., 12, 1087-1117, https://doi.org/10.5194/gmd-12-1087-2019

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., and Takahashi, K.: The JRA-55 Reanalysis: General Specifications and Basic Characteristics, 93, 5–48, https://doi.org/10.2151/jmsj.2015-001, 2015.

Kursa M., Rudnicki W., "Feature Selection with the Boruta Package" Journal of Statistical Software, Vol. 36, Issue 11, Sep 2010

Landsea, C. W. and Knaff, J. A.: How Much Skill Was There in Forecasting the Very Strong 1997–98 El Niño?, 81, 2107–2120, https://doi.org/10.1175/1520-0477(2000)081<2107:HMSWTI>2.3.CO;2, 2000.

Lee, Doo Young, Joong-Bae Ahn, Karumuri Ashok, and Andrea Alessandri. 2013. “Improvement of Grand Multi-Model Ensemble Prediction Skills for the Coupled Models of APCC/ENSEMBLES Using a Climate Filter.” Atmospheric Science Letters 14 (3): 139–45. https://doi.org/10.1002/asl2.430.

Lenssen, N. J. L., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., and Zyss, D.: Improvements in the GISTEMP Uncertainty Model, 124, 6307–6326, https://doi.org/10.1029/2018JD029522, 2019.

Liu, Q., L. Wang, Y. Qu, N. Liu, S. Liu, H. Tang, and S. Liang, 2013: Preliminary evaluation of the long-term glass albedo product. International Journal of Digital Earth, 6 (sup1), 69–95, doi:10.1080/17538947.2013.804601, https://doi.org/10.1080/17538947.2013.804601

Mishra, Niti, Chloé Prodhomme, and Virginie Guemas. 2018. "Multi-Model Skill Assessment of Seasonal Temperature and Precipitation Forecasts over Europe." Climate Dynamics 52 (7–8): 4207–25. https://doi.org/10.1007/s00382-018-4404-z.

Murphy, Allan H. 1973. "A New Vector Partition of the Probability Score." Journal of Applied Meteorology 12 (4): 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2.

Polo, I., Martin-Rey, M., Rodriguez-Fonseca, B. *et al.* Processes in the Pacific La Niña onset triggered by the Atlantic Niño. *Clim Dyn* **44,** 115–131 (2015). https://doi.org/10.1007/s00382-014-2354-7

Qian, S., Chen, J., Li, X., Xu, C.-Y., Guo, S., Chen, H., and Wu, X.: Seasonal rainfall forecasting for the Yangtze River basin using statistical and dynamical models, 40, 361–377, https://doi.org/10.1002/joc.6216, 2020.

Rayner, N. A.; Parker, D. E.; Horton, E. B.; Folland, C. K.; Alexander, L. V.; Rowell, D. P.; Kent, E. C.; Kaplan, A. (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century J. Geophys. Res.Vol. 108, No. D14, 4407 10.1029/2002JD002670.

Rodriguez, D., de Voil, P., Hudson, D., Brown, J. N., Hayman, P., Marrou, H., and Meinke, H.: Predicting optimum crop designs using crop models and seasonal climate forecasts, 8, 2231, https://doi.org/10.1038/s41598-018-20628-2, 2018.

Rodríguez-Fonseca, B., Polo, I., García-Serrano, J., Losada, T., Mohino, E., Mechoso, C. R., and Kucharski, F. (2009), Are Atlantic Niños enhancing Pacific ENSO events in recent decades? *Geophys. Res. Lett.*, 36, L20705, doi:10.1029/2009GL040048.

S2S4E project (GA n°776787) Deliverable D4.4, Skill assessment and comparison of methods for sub-seasonal and seasonal forecast systems for the energy sector.

Scaife, A.A., Smith, D. A signal-to-noise paradox in climate science. *npj Clim Atmos Sci* **1,** 28 (2018). https://doi.org/10.1038/s41612-018-0038-4

Schepen, A., Wang, Q. J., and Robertson, D. E.: Combining the strengths of statistical and dynamical modeling approaches for forecasting Australian seasonal rainfall, 117, https://doi.org/10.1029/2012JD018011, 2012.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, Theor Appl Climatol, 115, 15–40, https://doi.org/10.1007/s00704-013-0860-x, 2014.

Smith, D.M., Scaife, A.A., Eade, R. *et al.* North Atlantic climate far more predictable than models imply. *Nature* **583,** 796–800 (2020). https://doi.org/10.1038/s41586-020-2525-0

Scaife, A.A., Smith, D. A signal-to-noise paradox in climate science. npj Clim Atmos Sci 1, 28 (2018). https://doi.org/10.1038/s41612-018-0038-4

Torralba, V., Doblas-Reyes, F. J., MacLeod, D., Christel, I., and Davis, M.: Seasonal Climate Prediction: A New Source of Information for the Management of Wind Energy Resources, 56, 1231–1247, https://doi.org/10.1175/JAMC-D-16-0204.1, 2017.

van der Ploeg, T., Austin, P. C., and Steyerberg, E. W.: Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Med Res Methodol, 14, 137, https://doi.org/10.1186/1471-2288-14-137, 2014.

Weisheimer, A., and T. N. Palmer. 2014. "On the Reliability of Seasonal Climate Forecasts." Journal of The Royal Society Interface 11 (96): 20131162. https://doi.org/10.1098/rsif.2013.1162.

Wilks D.S., Statistical Methods in the Atmospheric Sciences, volume 100. Academic Press, 2011

Zhang, W., Villarini, G., Vecchi, G. A., Murakami, H., and Gudgel, R.: Statistical-dynamical seasonal forecast of western North Pacific and East Asia landfalling tropical cyclones using the high-resolution GFDL FLOR coupled model, 8, 538–565, https://doi.org/10.1002/2015MS000607, 2016.

The Added Value of Seasonal Climate Forecasting for Integrated Risk Management (SECLI-FIRM)

For more information visit

www.secli-firm.eu

or contact the SECLI-FIRM team at

info@secli-firm.eu

Grant Agreement n. 776868