S2S4E
Climate Services for Clean Energy

Research and Innovation action

H2020-SC5-2017

Skill assessment and comparison of methods for sub-seasonal and seasonal forecast systems for the energy sector

Deliverable D4.4

Version N°1

Authors: Ilias Pechlivanidis, Louise Crochemore (SMHI), Albert Soret, Llorenç Lledó, Andrea Manrique-Suñén (BSC), David Brayshaw, Paula Gonzlez, Andrew Charlton Perez, Hannah Bloomfield (UREAD), Franco Catalano, Irene Cionni (ENEA), Harilaos Loukos, Thomas Noël (TCDF)

GA n°776787

Disclaimer

The content of this deliverable reflects only the author's view. The European Commission is not responsible for any use that may be made of the information it contains.

Document Information

Grant Agreement	776787
Project Title	Sub-seasonal to Seasonal climate forecasting for Energy
Project Acronym	S2S4E
Project Start Date	01/12/2017
Related work package	WP 4. S2S Climate predictions
Related task(s)	Task 4.2: Comprehensive forecast quality assessment
Lead Organisation	SMHI
Submission date	01/03/2020
Dissemination Level	PU

History

Date	Submitted by	Reviewed by	Version (Notes)
21/01/2020	Ilias Pechlivanidis (SMHI)	All – SMHI, ENEA, UREAD, BSC	0.1
18/02/2020	Ilias Pechlivanidis (SMHI)	-	1.0

Table of content

About S2S4E 11
Summary 12
Keywords 13
Glossary 14
1. Introduction 16
2. Data and Methods 18
2.1. The forecasting systems 18
2.1.1. Seasonal time-scale 18
2.1.2. Sub-seasonal time-scale 18
2.1.3. Forecast horizons 19
2.2. The hydrological model 19
2.3. Demand and wind-power models 20
3. Improving skill – Use of pattern-based techniques 21
3.1. Highlights from deliverables D4.2 and D4.3 21
3.2. Discussion and Conclusions 23
4. Improving skill – Lagged ensembles and calibration 24
4.1. Skill assessment of NCEP CFS v2- lagged ensembles 24
4.2. Calibration and length of hindcast 26
4.3. Calibration window 27
4.4. Conclusions 29
5. Improving skill – Recommended Enhanced bias adjustment 31
5.1. Introduction 31
5.2. The reference methods 31
5.3. The CDFt method 32
5.4. Results 32
5.5. Discussion and Conclusions 37
6. Improving skill – Towards a multi-model approach 38
6.1. Introduction 38
6.2. Solar radiation and temperature 39
6.2.1. Introduction 39
6.2.2. Process-based model inter-comparison 40

6.2.3. Probabilistic scores and model independence 44
6.2.4. Discussion and Conclusions 47
6.3. A hydrological investigation 47
6.3.1. Methodology 47
6.3.2. Results 48
6.3.3. Discussion 52
6.3.4. Conclusions 53
6.4. Energy country average 53
6.4.1. Introduction and Motivations 53
6.4.2. Methodology 54
6.4.3. Results 55
6.4.4. Discussion and Conclusions 60
7. Improving skill – Seamless S2S forecasting 62
7.1. Introduction 62
7.2. Methodology 62
7.2.1. Forcing post-processing 62
7.2.2. Hydrological model runs 62
7.2.3. Forecast evaluation 63
7.2.4. Optimal combination horizon 63
7.3. A hydrological investigation – identification of critical lead times for S2S over Europe 63
7.3.1. When is the optimal combination horizon? 63
7.3.2. How does the optimal combination horizon vary spatially? 65
7.3.3. Where does the additional skill of sub-seasonal forecasts come from? 66
7.4. Discussion and Conclusions 67
8. Conclusions 70
Bibliography 72
Annex 1 - The impacts of using midnight vs. six hourly wind speeds to create wind power capacity factors for the S2S4E Decision Support Tool 77
Annex 2 - Improving skill – Recommended Enhanced bias adjustment 86
Annex 3 - Improving skill – Towards a multi-model approach – A hydrological investigation 90

List of figures

Figure 1. NCEP CFSv2 forecast setup. In hindcast mode each blue line is one ensemble member while in forecast mode each blue line corresponds to 4 ensemble members. ...24

Figure 2. 2m temperature Fair RPSS for tercile categories for NCEP CFSv2 for (rows) lagged ensemble of 4 members, lagged ensemble of 8 members, lagged ensemble of 12 members and for each of the 4 forecast weeks (columns). Hindcast period 1999-2010, reference ERA-Interim. 26

Figure 3. Differences in 2m temperature Fair RPSS between using a lagged ensemble of 8 members minus a lagged ensemble of 4 members (top row) and using a lagged ensemble of 12 members minus a lagged ensemble of 4 members for each of the 4 forecast weeks (columns). Hindcast period 1999-2010, reference ERA-Interim. 26

Figure 4. 2m temperature Fair RPSS for tercile categories for NCEP CFSv2 (12 members lagged ensemble) for raw forecast in the top row, NCEP CFS v2 calibrated with variance inflation using 12 years of hindcast (middle row) and NCEP CFS v2 calibrated with variance inflation using 'extended hindcast' of 20 years. The reference is ERA5. 27

Figure 5. 2m temperature Fair RPSS for tercile categories for ECMWF monthly forecast for January and 4 different lead times. The first row shows results for the raw forecast (raw), second row calibration has been applied with a weekly climatology (cal 1), in the third row calibration has been applied with a climatology based on 2 week window (3 start dates- cal 3), in the fourth row calibration has been applied with a climatology based on 3 week window (5 start dates- cal 5) and in the fifth row calibration has been applied with a climatology based on 4 week window (7 start dates-cal 7)....28

Figure 6. Similar with Figure 5 but for April....29

Figure 7. Annual CRPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the raw forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt). 34

Figure 8. Annual RPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt). 34

Figure 9. Annual CRPSS score of daily averages for precipitation for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt). 35

Figure 10. Annual RPSS score of daily averages for precipitation for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt). 35

Figure 11. a) 1-month-lead boreal winter (DJF) 2m temperature ECMWF minus Meteo 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

Meteo 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).41

Figure 12. 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 (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). 42

Figure 13. a) 1-month-lead boreal summer (JJA) surface solar radiation downward ECMWF minus Meteo 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 Meteo France for the predictions averaged over the Central-European domain (0E–16E; 45N–50N) 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). 43

Figure 14. 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 (0E–16E; 45N–50N) 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).....44

Figure 15. 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 Meteo 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 Meteo France; (d) ECMWF vs DWD...45

Figure 16. 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 Meteo 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 Meteo France; (d) ECMWF vs DWD. 46

Figure 17. Lead month 0 monthly spatial variability of the CRPSS of streamflow for the winter months..... 50

Figure 18. Lead month 0 monthly spatial variability of the CRPSS of streamflow for the summer months..... 51

Figure 19. CRPSS values of individual systems (MF, GLOSEA5 and SEAS5) and the multi-model (EMA) seasonal streamflow forecasts for all European sub-basins. The boxplots and outliers are based on results from all 35408 sub-basins in the E-HYPE model setup. 52

Figure 20. a) UK demand Q50 average pinball loss associated to the aggregation rules, the individual experts and the reference forecasts. The average losses for weeks 1 to 4 are presented as different symbols. The items on the x-axis are sorted from smaller to bigger loss based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for climatology, as a reference. b) Pinball losses for a subset of the models and references, but expressed as relative losses with respect to the UNIF_NWP combination, for each corresponding week. 56

Figure 21. Time-average weights involved in the Q50 aggregations: BOA (black bars, top panels), MLpol (black bars, bottom panels), BOA_NWP (grey bars, top panels) and MLpol_NWP (grey bars, bottom panels); for leads: a) week1, b) week2, c) week3 and d) week4. 57

Figure 22. a) Temporal evolution of UK demand for week 4. The black line corresponds to reanalysis, the grey line corresponds to the leave-one-out climatology and the red and blue lines are the ensemble means for ECMWF and NCEP, respectively. Weight evolutions for the aggregation rules corresponding to that same lead time for a) BOA and b) MLpol. 58

Figure 23. a) UK demand quantile-mean average pinball loss associated to the aggregation rules, the individual experts and the reference forecasts. The average losses for weeks 1 to 4 are presented as different symbols. The items on the x-axis are sorted from smaller to bigger loss based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for climatology, as a reference. b) Pinball losses for a subset of the models and references, but expressed as relative losses with respect to the UNIF_NWP combination, for each corresponding week. 60

Figure 24. Median CRPSS in terms of streamflow (top) and snow water equivalent (bottom) over Europe per forecast initialisation date. Each colour corresponds to a calendar month. Thick lines correspond to the seasonal forecasts and thinner lines correspond to the sub-seasonal forecasts. 64

Figure 25. Maps of optimal combination horizons for streamflow (top) and snow water equivalent (bottom). Each column corresponds to a different sub-seasonal forecast issue week within the month..... 66

Figure 26. Maps of optimal combination horizons for precipitation (top) and temperature (bottom). Each column corresponds to a different sub-seasonal forecast issue week within the month. 67

Figure 27. Skill regions identified over Europe for (a) streamflow and (b) snow water equivalent (right). Regions where sub-seasonal forecasts have a longer combination horizon appear in shades of red, others appear in grey. 69

List of tables

Table 1. Technical details of the seasonal prediction systems. 18

Table 2. Summary of the methodological details of the sub-seasonal and seasonal approaches (used in D4.2 and D4.3). 21

Table 3. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for 2m temperature daily averages for the first three lead times of the uncorrected forecast and the forecast adjusted daily with the three different methods spatially averaged over Europe, Africa, East-Asia and North America. 36

Table 4. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for precipitation averages for the first three lead times of the uncorrected forecast and the forecast adjusted daily with the three different methods spatially averaged over Europe, Africa, East-Asia and North America. 37

Table 5. Model configuration, resolution, and initialization strategy of each contributing system. 39

About S2S4E

The project seeks to improve renewable energy variability management by developing a tool that for the first time integrates sub-seasonal to seasonal climate predictions with renewable energy production and electricity demand. Our long-term goal is to make the European energy sector more resilient to climate variability and extreme events. Large-scale deployment of renewable energy is key to comply with the emissions reductions agreed upon in the Paris Agreement. However, despite being cost competitive in many settings, renewable energy diffusion remains limited largely due to seasonal variability. Knowledge of power output and demand forecasting beyond a few days remains poor, creating a major barrier to renewable energy integration in electricity networks. To help solve this problem, S2S4E is developing an innovative service to improve renewable energy variability management. The outcome will be new research methods exploring the frontiers of weather conditions for future weeks and months and a decision support tool for the renewable industry.

More information: www.s2s4e.eu

Summary

Renewable energy (i.e. wind, solar, hydro, electricity demand) is the fastest growing source of electricity globally. Although renewable energy diffusion still faces important challenges related to large-scale integration in the energy system, there is already high potential over the European energy market. The S2S4E project aims to: 1) foster renewable energy deployment while also maintaining energy security by providing sub-seasonal to seasonal (S2S) climate forecasts; 2) enable the energy industry and policy makers to assess how well different renewable energy sources will meet demand over extended time horizons (weeks to months), focusing on the impact of climate variables on energy outputs and needs; 3)make the European energy sector more resilient to climate variability and high impact events by exploring the frontiers of what can be achieved by using S2S predictions offering a new decision support tool based on S2S climate predictions; 4) contribute for the expansion of climate services to users and markets as it will base its development on a user-centric framework for co-design and co-development.

To achieve the projects objectives, an exploration of the scientific frontiers (multi-modelling, downscaling, bias-adjustment etc.) of sub-seasonal predictions and their synergies with seasonal predictions is needed, allowing the provision of useful information for decision-making process. In addition, statistical assessment of the predictions and their added-value from S2S predictions based on a historical period and historical case studies is important for improved communication with the users of the energy sector. In the setup of a forecasting service, a benchmarking assessment is necessary, in which a single model approach (i.e. analyzing forecasts of individual climate systems form different data/service providers) is employed, based also on simple bias-adjustment methods to downscale the forecasted information at the local scale. The outcomes of this benchmarking assessment allow the users to have a first vision of the potential application of S2S climate predictions for decision-making. However, a number of different pre- or post-processing methods/approaches have been proposed to achieve a service evolution.

This report builds upon the need to assess state-of-the-art methods and approaches that have shown to provide improved predictions for the energy sector. The report is also building on the conclusions from deliverables D4.2 and D4.3 which quantify the impact of weather regimes on the energy sector and the use of teleconnections to construct energy-relevant variables. In addition, assessments of methodologies including calibration of lagged ensembles, bias-adjustment, multi-modeling and seamless forecasting, are performed. These are done for different essential climate variables and indicators relevant to the energy sector. The assessments are conducted at two different temporal horizons, i.e. sub-seasonal and seasonal, with the forecast information being provided by different providers. Some of the assessments are done over the European or global domain, which complements the “deep” knowledge from regional scale assessment, whilst giving an overview of the improved predictability to users with large-scale interests. Large-scale assessment has the potential to encompass many regions, cross-regional and international boundaries and represent a number of different climatic zones and hence it allows for exploration of emerging patterns and facilitation of comparative analysis.

Keywords

Forecasting; Sub-seasonal to Seasonal; Energy sector; Climate services; Multi-modelling; Pattern-based analysis; Lagged ensembles; Bias-adjustment; Seamless forecasting

Glossary

Bias-Adjustment (BA)	Process aiming at removing systematic errors in the output of a model. Methods include: linear scaling, distribution-based scaling, quantile mapping.
Downscaling	General name for a procedure to take information known at large scales to make predictions at local scales. The two main approaches to downscaling climate information are dynamical and statistical. Statistical methods include: linear scaling, distribution-based scaling, quantile mapping.
ECMWF	European Centre for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
ENS-ER	ECMWF Extended Range (sub-seasonal forecasts) ensemble prediction system
Forecast time	The time between the initiation and completion of a forecast.
Forecast quality	How well a forecast compares against a corresponding observation of what actually occurred, or some good estimate of the true outcome.
GloSEA5	UK MetOffice long-range System 13 (seasonal forecasts)
Initial conditions (IC)	The hydrological states (soil moisture, snow cover, water already in the river, among others) at or close to the start of the forecast run.
MF	Meteo France long-range System 6 (seasonal forecasts)
MME	Multi Model Ensemble, combination of forecasts from different forecast systems
Seasonal climate forcing (SCF)	The seasonal meteorological forecast used as input to a hydrological model.

S2S	Sub-seasonal to seasonal forecasts
S2S4E	The project "Sub-seasonal to seasonal climate predictions for energy"
SEAS5	ECMWF long-range System 5 (seasonal forecasts)

1. Introduction

Climatic information from sub-seasonal (6 weeks ahead) to seasonal (6 months ahead) time-scales is needed for decision-making in a number of sectors. Compared to the short-to-medium-range (up to 10 days ahead) forecasts, S2S time-scales hold the potential for being of great value for a wide range of users who are affected by variability in climate, water and energy and who would benefit from understanding and better managing climate-related risks (Bruno Soares et al., 2017; Stoft, 2002; Green, 2005). Wind, solar and hydro and electricity demand are examples of energy applications in which S2S information can affect decision-making.

In Europe, there has been relatively little uptake and use of S2S forecasts by users for decision making, compared to other parts of the world, such as the USA and Australia, probably due to the relatively limited inherent predictability and limited quality of models and observations (Bennett et al., 2017; Mendoza et al., 2017; Arnal et al., 2018). However, recent advances in our understanding and forecasting of climate have resulted in skillful and useful climatic predictions, which can consequently increase the confidence of energy-related predictions, and improve awareness, preparedness and decision-making from a user perspective (Bruno Soares and Dessai, 2016).

The accuracy of S2S energy forecasts is subject to multiple sources of error/uncertainty, which are present in the various components of the production chain, i.e. initialization and production of the climate forecasts, bias-adjustment, impact model(s) (Crochemore et al., 2016; Demirel et al., 2013; Thiboultet al., 2016). Consequently, to improve forecasts, each component has to be evaluated to assess its relevant contribution to the overall forecasting accuracy (Arnal et al., 2017; Wood and Lettenmaier, 2008; Yossef et al., 2017). In addition, predictability is characterized by strong spatial variation and commonly a temporal degradation of its skill in longer time-scales (Arnal et al., 2017; Greuell et al., 2018). The large heterogeneity in the climatic patterns and physiographic descriptors results in a strong spatiotemporal variability of the predictability. Furthermore, the understanding of the key drivers influencing predictability is still limited.

In order to address the scientific and technical challenges for improved climate services, S2S4E aims to: 1) foster renewable energy deployment while also maintaining energy security by providing sub-seasonal to seasonal (S2S) climate forecasts; 2) enable the energy industry and policy makers to assess how well different renewable energy sources will meet demand over extended time horizons (weeks to months), focusing on the impact of climate variables on energy outputs and needs; 3) make the European energy sector more resilient to climate variability and high impact events by exploring the frontiers of what can be achieved by using S2S predictions offering a new decision support tool based on S2S climate predictions; 4) contribute to the expansion of climate services to users and markets as it will base its development on a user-centric framework for co-design, co-development and co-evaluation.

Within the S2S4E project, a first assessment of the forecast skill of sub-seasonal and seasonal forecast systems applied to energy was provided in the deliverable D4.1. That deliverable focused on evaluating skill in directly forecasting surface meteorological variables (e.g., wind, temperature, precipitation, and insolation; often referred to as Essential Climate Variables or ECVs) and their subsequent conversion into energy-relevant quantities (wind power, demand,

hydrology, solar power). In general, skill was shown to exist over some regions of Europe - but at rather modest levels - for multi-week (sub-seasonal) and multi-month (seasonal) lead times.

This report collects the scientific effort within the S2S4E project to improve the forecasting skill for different geographical domains, variables of interests, and seasons. The investigations focus on two time horizons, i.e. sub-seasonal and seasonal. A number of different state-of-the-art approaches and methodologies that could potentially improve the forecasting skill are presented. The methods include: patterns-based techniques, calibration methods for lagged ensembles, bias-adjustment methods, multiple models and methods to average them, and seamless S2S forecasting. Those methods have showed to be on the front scientific line with promised potential based on the S2S4E partners' experience.

A background of the S2S forecasting systems and the impact models used is given in Section 2. An assessment of the use of pattern based techniques for improving forecasts is given in Section 3, followed by an assessment of the lagging ensemble including their calibration in Section 4. Section 5 assesses the impact of bias-adjustment and the methodology adapted on the different forecasts. Section 6 presents the added value on forecasting predictability by introducing information from multiple climate models. An assessment of seamless sub-seasonal to seasonal forecasting for hydrology is presented in Section 7. Finally, Section 8 states the conclusions.

2. Data and Methods

Many of the datasets and tools used in this document follow closely the methods developed in earlier deliverables (particularly D3.1, D3.2 and D4.1) and in the "partner" deliverables D4.2 and D4.3. As these datasets and tools are central to the science that follows, the following sections seek to provide a high-level overview of the datasets and methods involved. More comprehensive discussion of each dataset/tool can be found in previous documentation (references provided).

Consistent with the research objectives of this deliverable, however, it is noted that different research activities (generally corresponding to the individual chapters) have introduced a range of experimental innovations to the basic techniques in order to advance understanding and/or improve predictive skill. As such, the implementation of each dataset, tool or method for a particular research task is provided separately within each of chapters 3 to 7.

2.1. The forecasting systems

2.1.1. Seasonal time-scale

Several European national meteorological centers and institutions produce operational seasonal predictions. Seven different seasonal prediction systems are available from from the Climate Data Store (CDS) of the Copernicus Climate Change Service (C3S) initiative: European Center for Medium-Range Weather Forecasts (ECMWF), Deutscher Wetterdienst (DWD), Meteo France (MF), UK Met Office (UKMO) and Centro Euro- Mediterraneo sui Cambiamenti Climatici (CMCC). , The C3S service provides a unified access point, and a common hindcast period and spatial resolution. Some details of each of the prediction systems employed here, as the number of ensemble members, the hindcast period or the spatial grid are detailed in Table 1.

Center	Prediction system	Analyzed period	Ensemble members	Horizontal grid
CMCC	SPS3	1993-2018	40	Regular 360x180
DWD	System2	1993-2018	30	Regular 360x180
UKMO	GLOSEA5 (System13)	1993-2018	28	Regular 360x180
MF	System6	1993-2018	25	Regular 360x180
ECMWF	SEAS5	1993-2018	25	Regular 360x180

Table 1. Technical details of the seasonal prediction systems.

2.1.2. Sub-seasonal time-scale

In the sub-seasonal time range, two systems from the S2S database (Vitart et al., 2017) were employed: ECMWF monthly forecast system (MFS or extended range) and NCEP CFSv2.

ECMWF-MFS (Vitart, 2004) runs coupled ocean-atmosphere integrations up to 46 days issued every Monday and Thursday. Operational configuration consists of 51 ensemble members while the hindcasts consist of 11 members. It was described in deliverable D4.1 (Section 3.1.2) and its skill for surface variables over Europe was assessed in deliverable D4.1 (Section 5). The National Centers for Environmental Prediction's (NCEP) Climate Forecast System (Saha et al., 2014) is a coupled system to both an ocean model (GFDL MOM4) and an ice model. The forecast length for sub-seasonal predictions is 45 days and the system is run every 6 hours. The real-time forecast runs three perturbed members and one control run initialized four times a day (00, 06, 12 and 18 UTC). The hindcast period is fixed and spans 12 years (1999-2010). The hindcast is also initialized daily, four times a day, but only one simulation at the time, producing a lagged ensemble of 4 members daily (in some cases, a larger ensemble is created by including a wider lagging period, e.g., up to 3 days previous).

2.1.3. Forecast horizons

In terms of the temporal scale analysis, forecast skill is evaluated using the following lead time convention. For seasonal forecasts, month 1 corresponds to one month after initialization (so for an ensemble of forecasts launched at any point in November, forecast month 1 is December). For sub-seasonal forecasts, week 1 is defined as the week starting at day 5 (i.e. forecast week 1 is the period day 5 to 11, week 2 is days 12-18 etc.). A complication occurs in the case of lagged ensemble forecasts, which may include ensemble members launched earlier (in which case day 5 is defined relative to the most recently launched ensemble member).

2.2. The hydrological model

For the investigations that require hydrological modelling, e.g. streamflow and snow, we are benefited from a setup of the HYPE hydrological model at the continental scale. The HYPE model has been setup for the pan-European region (8.8 million km^2) (Hundecha et al., 2016) with a spatial resolution of about 35400, i.e. in average 215 km^2, and is referred to as E-HYPE v3.0. The HydroGFD meteorological forcing dataset (daily mean precipitation and temperature), an observation corrected reanalysis (Berg et al., 2018) for the period 1981 – Today, is used as reference for model calibration and verification.

The model was calibrated to secure usefulness to the potential users and applications; hence an adequate model performance in terms of discharge and other hydrological variables is important. Parameters are linked to catchment descriptors with good transferability, with median Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970) of 0.54 and 0.53, and median volume error of −1.6% and 1.3% in the calibration and verification stations, respectively. Note that in this investigation, bias-adjusted seasonal meteorological forecasts from ECMWF ENS-ER and SEAS5 are used to force the E-HYPE hydrological model. Further information about the hydrological model setup over the pan-European domain can be found in deliverable D4.1 (section 3.2).

2.3. Demand and wind-power models

A brief overview of the country aggregate demand and wind power capacity factor models is given below. Full details of the country aggregate models can be found in Bloomfield et al. (2019) and S2S4E Deliverable 3.2 Annex A.

Electricity demand is calculated with a country-level multiple-linear regression model containing parameters to capture both meteorological and human behavior. Each country has a unique regression model, which is trained on two years of measured demand data (2016-2017) from the ENTSOe transparency platform (ENTOSe, 2019), and is then applied retrospectively to the full ERA5 reanalysis period (1980-2018). Two versions of the model output are created, the "full" demand (using all of the available regression parameters) and the "weather-dependent" demand (which includes only the weather-dependent terms, heating-degree-days and cooling-degree-days – i.e., removes the impacts of the day-of-week behavioral patterns and long term socio-economic trends). In this chapter the weather-dependent model is used to highlight the meteorologically driven power system variability.

Wind power capacity factor is calculated based on the methodology of Lledó, et al. (2017) and Lledó et al. (2019) which calculates gridded capacity factor using three different power curves corresponding to three turbine classes. To calculate country aggregate capacity factor, firstly the most appropriate wind turbine for each grid box is calculated based on the 1980-2018 mean ERA5 100m wind speed. Previous work in Deliverable 4.1 highlighted that ERA5 produces anomalously low 100m wind speeds over large regions of Europe; therefore prior to use within the wind power model the ERA5 100m wind speeds are bias corrected to the global wind atlas (Global Wind Atlas, 2019). The country aggregate capacity factor is calculated by passing bias corrected 100m wind speeds through each curve and aggregating based on the locations in installed turbines taken from the windpower.net (2019).

As stated in section 2.2, within the S2S hindcasts meteorological variables are only available once daily. For 2m temperature (the input for the demand model) daily average temperatures are available, which is the same input used in the ERA5 demand model. However, for the wind power model, midnight 10m wind speeds are available from the hindcasts, as opposed to six-hourly 100m wind speeds in ERA5. The potential impacts of this on the wind power model performance are discussed in Annex 1.

3. Improving skill – Use of pattern-based techniques

3.1. Highlights from deliverables D4.2 and D4.3

The skill of seasonal and sub-seasonal prediction systems in forecasting surface conditions that impact the energy sector was analyzed at grid-point level in the deliverable D4.1. The results highlighted some windows of opportunity to employ those forecasts of surface conditions and of derived generation/demand indicators over Europe. However, in most cases there is room for improvement. The dynamical prediction systems are known to have biases in the shape, magnitude and location of the atmospheric circulation patterns, compared to observations. Also, the relationship between large-scale circulation and surface conditions or energy variables is not always well represented in those prediction systems. Although bias adjustment techniques have been investigated during the project and applied thoroughly in the Decision Support Tool (DST) to address those issues, in deliverables D4.2 and D4.3 a slightly different approach was investigated to produce better predictions. Several specific methods have been proposed, with the overarching idea that large-scale circulation patterns might be more predictable than small-scale features, and also that defective relationships between the circulation patterns and energy variables in the predictions can be replaced by more accurate observed relationships employing hybrid dynamical-statistical methods.

In deliverable D4.2 several machine learning techniques (such as k-mean clustering, fuzzy sets, analogues or principal component analysis) have been employed to analyze and classify large-scale circulation fields. Then the ability of seasonal and sub-seasonal predictions to anticipate the occurrence of those circulation patterns has been analyzed. Furthermore, in deliverable D4.3 those forecasts of large-scale patterns have been employed to derive forecasts of surface variables or energy indicators. In short, those methods boil down to:

a) Produce a set of circulation patterns from historical observations
b) Predict the occurrence of those patterns from dynamical prediction systems
c) Employ past observed relationships between patterns and energy variables/indicators to reconstruct forecasts of those variables/indicators.

This type of approach is also known as Perfect Prognosis (or PerfectProg) in short-term weather forecasting, because one assumes a perfect forecast of an intermediate variable and derives an impact forecast based on past observations, i.e. there is no intent to correct the biases of the model in predicting the intermediate variable. Table 2 summarizes the methodological details of the different approaches presented in D4.2 and D4.3.

These methods have proven useful for some lead times, regions and periods of the year and can improve the skill of dynamical predictions at producing energy-relevant variables or demand/generation indicators. Additionally, the provision of circulation pattern forecasts and forensic analyses can also be directly of interest to many climate service users. Therefore, it could be useful to implement those methodologies operationally to produce enhanced products.

	Method	Pattern recognition algorithm	Method to produce pattern forecasts	Forecasted circulation patterns	Statistical reconstruction of energy variables	Forecasted energy variables	Analyzed periods
Seasonal	Euro-Atlantic teleconnections	Rotated Empirical Orthogonal Functions of Z500 seasonal anomalies	Projection of seasonal-mean forecasts of Z500 onto observed teleconnection patterns	4 teleconnection indices	Multilinear regression	grid-point fcsts of wind, temperature & solar radiation.	DJF/MAM/JJA/SON
Seasonal	Weather Regimes	K-mean clustering of daily MSLP patterns	Assignation of daily MSLP forecasts to each regime with minimum distance	Frequency of occurrence of the 4 weather regimes	Weighted mean of centroid winds	grid-point fcsts of surface wind	All months
Seasonal	Hydrological Weather Regimes	Fuzzy sets of daily MSLP patterns	Assignation of daily MSLP forecasts according to the degree of fulfilment (DOF) for each fuzzy rule	Frequency of occurrence of the 12 fuzzy sets	Analogues of temperature and precipitation + HBV-96 hydrological model	Inflows for Ume river	All months
Sub-seasonal	Weather Regimes	K-mean clustering of daily MSLP patterns	Assignation of daily MSLP forecasts to each regime with minimum distance	Frequency of occurrence of the 4 weather regimes			All 52 weeks of the year
Sub-seasonal	Targeted Circulation Types	K-mean clustering of country-aggregate daily demand-net-wind	Assignation of daily Z500 ensemble-mean forecasts to each regime with minimum distance to Z500 centroid	Daily evolution of the 4 targeted circulation types	Weighted mean of demand-net-wind centroids or Demand-net-wind centroid of dominant targeted circulation type	Country-aggregates of energy demand-net-wind	All weeks from Nov to Mar

Table 2. Summary of the methodological details of the sub-seasonal and seasonal approaches (used in D4.2 and D4.3).

3.2. Discussion and Conclusions

Most of the scientific developments produced in D4.2/D4.3 are difficult to implement in the current structure of the S2S4E DST, which has been designed to present probabilistic forecasts of dynamical models (i.e. with ensemble members, tercile and decile probabilities, hindcasts, etc.). During its co-development, the tool was not intended to present forecasts of circulation indices or patterns. Also, presenting different methods in the tool simultaneously or changing existing procedures can be confusing to users and damage the project reputation. Therefore, caution has to be taken before implementing new methods in the DST. However, there are options to provide additional services through other channels with these techniques. For instance, forecasts of Euro-Atlantic Teleconnections can be very useful for seasonal forecast outlooks (indeed forecasts of NAO are routinely presented in some European winter climate outlooks already, but the same cannot be said for the other teleconnections and seasons).

4. Improving skill – Lagged ensembles and calibration

4.1. Skill assessment of NCEP CFS v2- lagged ensembles

The forecast system providing the sub-seasonal forecasts in the first version of S2S4E Decision Support Tool (operational since June 2019) is the NCEP Climate Forecast System (CFSv2, Saha et al., 2014), which also provides seasonal predictions in the C3S service. The NCEP CFSv2 system was chosen for operationalization because 1) it is freely available 2) it provides instantaneous values of the wind components, necessary to calculate a valid wind speed module. However, due to the model configuration, the use of NCEP CFSv2 poses some difficulties related to the lagged initializations and the short hindcast size.

NCEP CFS v2 employs the NCEP GFS atmospheric model and the GFDL MOM4 ocean and sea ice model. The original horizontal resolution is T126 (100 km) and 64 levels in the vertical for the atmosphere while the ocean model has a horizontal resolution of 0.5° in the zonal direction and 0.25° in the meridional direction with 40 vertical layers. In contrast to other sub-seasonal or seasonal prediction systems that produce ensembles in "burst" mode (e.g. ECMWF monthly prediction system), NCEP CFSv2 is run on a sub-daily basis producing multiple forecast runs that can be combined in a lagged ensemble. Forecast runs are initialized every 6 hours (00 UTC, 06 UTZ, 12 UTC and 18 UTC) with 4 ensemble members (3 perturbed members and one control run). The forecast length of each of these runs is at least 45 days. The hindcast period is fixed and spans 12 years (1999-2010). The hindcast is also initialised every 6 hours, four times per day but with only one member. Overall, in forecast mode there are 16 initialisations per day while in hindcast mode there are 4 initializations per day. This setup is illustrated in Figure 1.

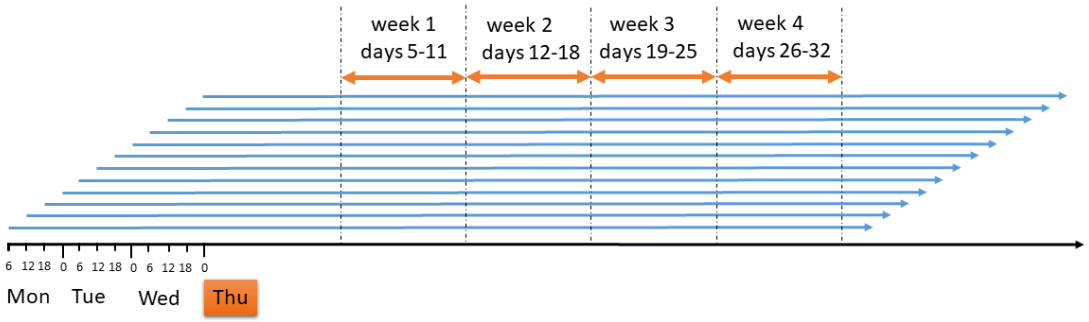


Figure 1. NCEP CFSv2 forecast setup. In hindcast mode each blue line is one ensemble member while in forecast mode each blue line corresponds to 4 ensemble members.

With this type of setup a lagged ensemble can be built to increase the number of ensemble members of a forecast. A lagged ensemble is constructed by pooling together forecasts from the same model initialised at different times. For example, the forecast initialised at 00:00 Thursday for the days 5-11 (following Monday-Sunday period) is composed of 4 ensemble members. If the ensemble members corresponding to the previous model run (18:00 Wednesday) verifying at the same week are added, the number of ensemble members would increase to 8; if also the ensemble members corresponding to the Wednesday 12:00

initialisation are added, then the total of ensemble members would be 12, and so on. Increasing ensemble size can increase skill by increasing spread, but it can also degrade skill by including older information from longer lead times. Determining the optimal lagged ensemble is not trivial as it depends on the skill score considered, the variable and the lead time. Moreover, these analyses are often performed on the hindcast dataset with less ensemble members as the hindcast offers a longer sample to compute skill scores. A study of optimal lagged ensemble of NCEP CFSv2 to predict MJO found that for lead times shorter than 7 days, adding more than one initialisation degraded skill, and for larger than 7 days approximately 5 lagged members were optimal (Trenary et al. 2017). However, this study employed a deterministic score, the mean square forecast error. Other studies that have assessed the forecast skill of NCEP CFSv2 have used 4-day lagged ensemble (16 members) in the hindcast (DelSole et al., 2017) and 3-day lagged ensemble (12 members) in the hindcast (Wang and Robertson, 2019).

To assess the impact of the number of ensemble members on the fair RPSS for tercile categories several tests were conducted using the hindcast period 1999-2010. As a first test 4 ensemble members were aggregated (1-day lagged ensemble), then 8 ensemble members (2-day lagged ensemble) and 12 ensemble members (3-day lagged ensemble). Maps of fair RPSS for tercile categories of 2m forecasts with the lagged ensembles are shown in Figure 2 for each of the 4 forecast weeks: week 1 (days5-11), week 2 (days 12-18) week 3 (days 19-25) and week 4 (days 26-32) for January. The reference for the skill score is ERA-Interim reanalysis. Similar patterns of skill are identified in the 3 cases. Week 1 presents high skill for most regions. However, there is an area of low skill in the Maritime continent which is believed to be a consequence of the use of ERA-Interim as reference for NCEP CFSv2 (Kim et al. 2012). They demonstrated that the skill of NCEP CFSv2 was considerably better when compared against CFSR (Climate Forecast System reanalysis) than when compared against ERA-Interim, particularly in the tropical ocean. For weeks 2, 3 and 4 the skill scores diminishes considerably, although there are some regions that retain skill for longer lead times. The most skillful areas over ocean are the eastern equatorial Pacific (El Niño region) and the tropical Atlantic (north and south). Over land, the areas that show highest skill are the Amazon basin (up to week 4), North America and central Asia.

The differences in skill amongst the 3 lagged ensembles are too small to be seen in Figure 2. To better identify any improvement, the differences in skill between lagged ensemble 8 and 12 minus lagged ensemble 4 are shown in Figure 3. Using 8 members to produce a lagged ensemble has very little impact on the skill score while increasing the number of members to 12 produces a slight improvement, particularly for weeks 1 and 2 in the equatorial region, Atlantic, Indian Ocean.

A lagged ensemble over 3 days was considered a good compromise to be used in the DST. In the case of forecast, the total number of ensemble members composing a 3 days lagged ensemble is 48. Since the skill assessment is done with the hindcast, the actual skill corresponding to the real-time forecasts is expected to be higher because of the larger ensemble.

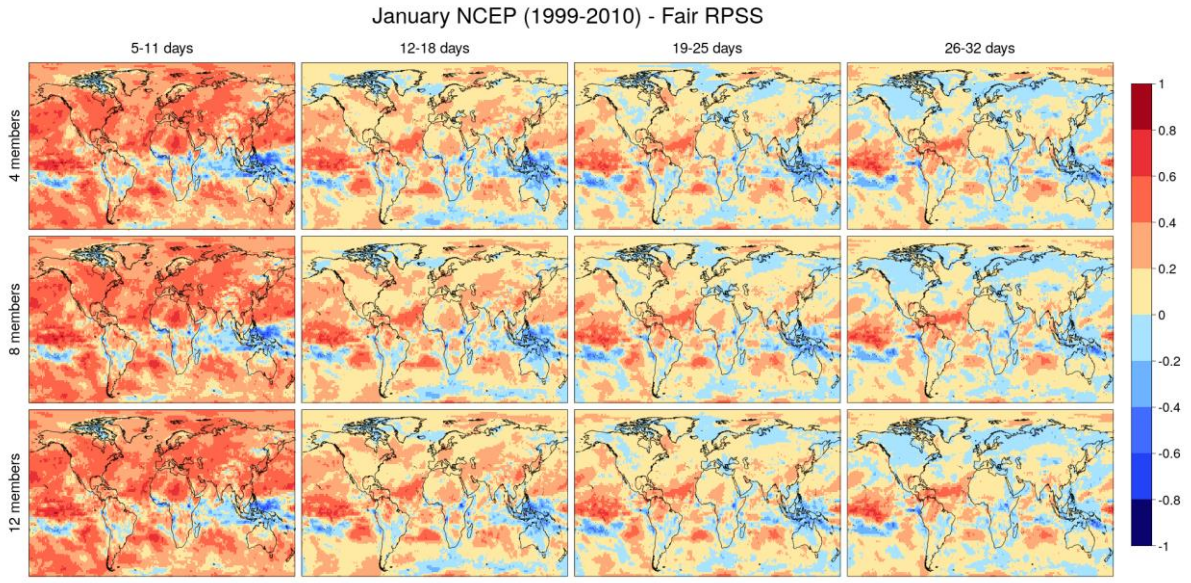


Figure 2. 2m temperature Fair RPSS for tercile categories for NCEP CFSv2 for (rows) lagged ensemble of 4 members, lagged ensemble of 8 members, lagged ensemble of 12 members and for each of the 4 forecast weeks (columns). Hindcast period 1999-2010, reference ERA-Interim.

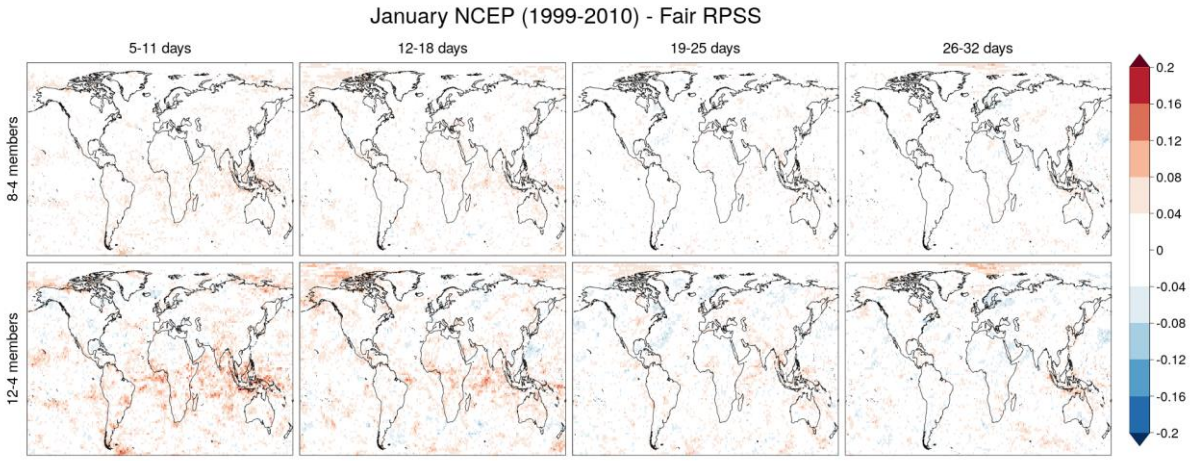


Figure 3. Differences in 2m temperature Fair RPSS between using a lagged ensemble of 8 members minus a lagged ensemble of 4 members (top row) and using a lagged ensemble of 12 members minus a lagged ensemble of 4 members for each of the 4 forecast weeks (columns). Hindcast period 1999-2010, reference ERA-Interim.

4.2. Calibration and length of hindcast

Sub-seasonal forecasts need to be post-processed in order to remove the model drift and adjust the members spread. The variance inflation calibration method (Doblas-Reyes et al., 2005; Torralba et al. 2017) has been used to calibrate seasonal forecasts (Manzanas et al., 2019; Torralba et al., 2017). For sub-seasonal forecast the implementation poses some challenges. In the case of NCEP CFSv2, the hindcast period is 1999-2010 spanning only 12 years. To increase the amount of years to be used as sample for the calibration, forecasts after 2010 have been

concatenated to the hindcast period. This way, the hindcast was extended by adding the control runs of every 6 hourly forecast initialisation from 2011 until 2018. Some months of data were missing so the calibration for each week was done with the available years. The calibration is applied to every forecast week independently.

Figure 4 shows some results of tests with the calibration, in this case all the forecasts are a 3-day lagged ensemble of 12 members. In the top row of Figure 4 the fair RPSS for tercile categories is shown, in this case the reference is ERA5 and the skill is calculated for the period 1999-2018. In this case no calibration is applied. The pattern is similar to the maps of Figure 2. The second row shows the fair RPSS when a calibration with a 12 years hindcast is applied. The method is able to reduce the negative skill of some areas like South America, South Africa, and the Maritime Continent (although in this region some negative skill still persists). For weeks 3 and 4, however, the skill over land in the north hemisphere is reduced. However in the equatorial region even over land there is skill up to week 4 (Amazon basin North Africa, South Asia). In the last row, the calibration has been performed using the extended hindcast with 20 years of data (1999-2018). It can be seen how the larger sample for calibration has a positive impact on skill. The improvement is particularly evident over the mid latitudes in both hemispheres, where areas that showed negative skill with the 12-year calibration now present positive skill.

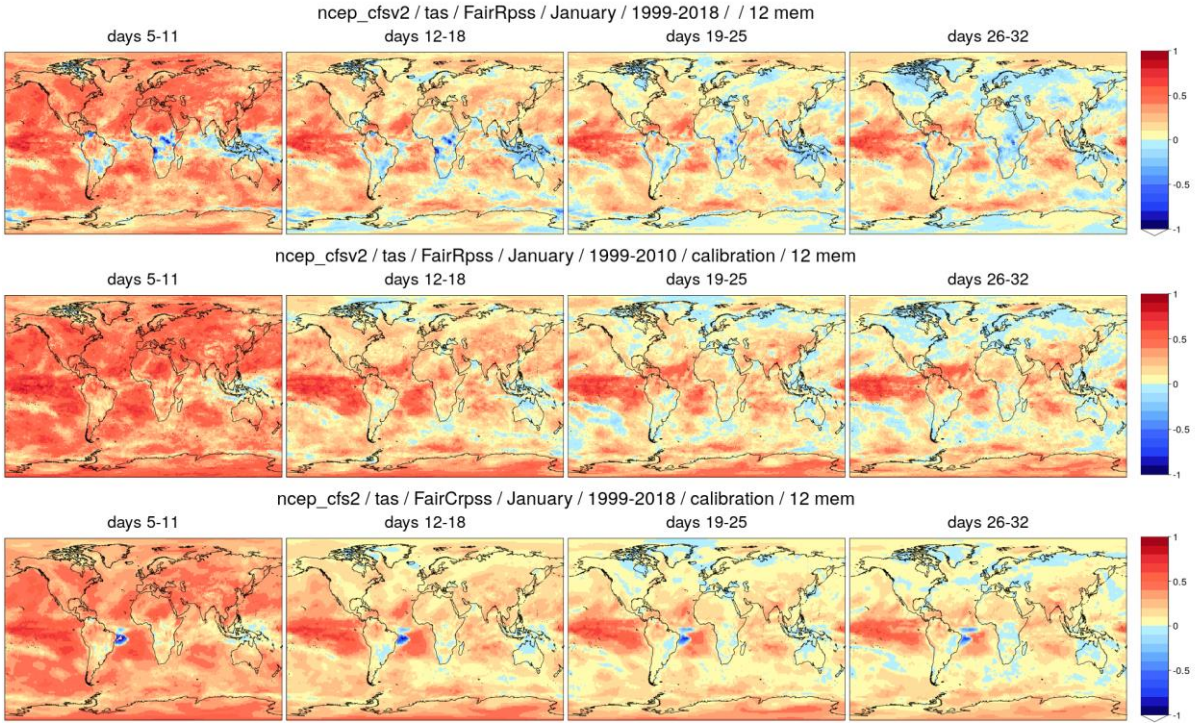


Figure 4. 2m temperature Fair RPSS for tercile categories for NCEP CFSv2 (12 members lagged ensemble) for raw forecast in the top row, NCEP CFS v2 calibrated with variance inflation using 12 years of hindcast (middle row) and NCEP CFS v2 calibrated with variance inflation using 'extended hindcast' of 20 years. The reference is ERA5.

4.3. Calibration window

Another option to increase the sample data for calibration is to use a window around the interest week. In this case forecasts are from ECMWF monthly system which produces a hindcast of 20 years for every forecast. Several options have been tested to determine the optimal length

of the window. The size of the window for calibration in an operational context is limited by the availability of the hindcasts corresponding to the real-time forecasts. ECMWF distributes the hindcast 2 weeks before each forecast, so for a forecast issued on a Thursday, the widest centered window from the hindcasts that can be used is of seven semi-weekly start dates (issued on Mondays and Thursdays). Taking into account this limitation, the window sizes tested are: 2-week window (one start date before and one after the start date being calibrated, total 3 start dates-cal 3), 3-week window (two start dates before and two after the start date being calibrated, total 5 start dates-cal 5) and 4-week window (three start dates before and three after the start date being calibrated, total 7 start dates-cal 7). As an illustration, for a forecast issued on Thursday, the 3 start dates (2 week window) imply using the hindcast corresponding to the previous Monday, the hindcast corresponding to that Thursday and the hindcast corresponding to the following Monday. The increase of the sample size produces a more robust climatology. However, increasing the window, information from weeks which are further away from the target week is used which can degrade the adjusted forecast. It should be noted that the model climatology is computed for each lead time independently.

Results are shown in Figure 5 for the month of January. The most interesting differences appear for weeks 3 and 4. It can be seen how a simple calibration using one week already reduces some areas of negative skill. When increasing the calibration window to 2 weeks, the skill becomes positive in almost all regions for week 4. Successive increments of the calibration window lead to same results (cal 5 and cal 7).

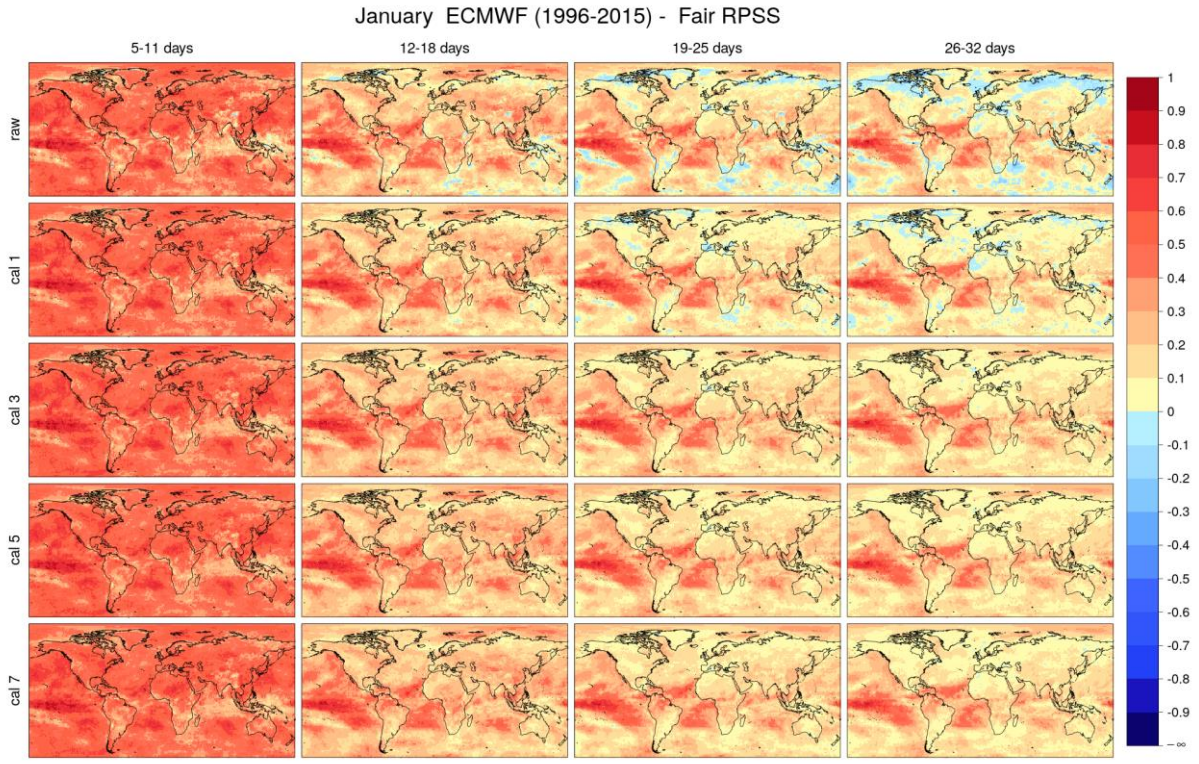


Figure 5. 2m temperature Fair RPSS for tercile categories for ECMWF monthly forecast for January and 4 different lead times. The first row shows results for the raw forecast (raw), second row calibration has been applied with a weekly climatology (cal 1), in the third row calibration has been applied with a climatology based on 2 week window (3 start dates- cal 3), in the fourth row calibration has been applied with a climatology based on 3

week window (5 start dates- cal 5) and in the fifth row calibration has been applied with a climatology based on 4 week window (7 start dates-cal 7).

The same analysis is conducted for April, results are shown in Figure 6. In this case, the increase in the window beyond 3 start dates results in a degradation of skill in some areas of central Asia. The reason behind is that because of the seasonal cycle, a longer window is not representative of the climatology of the week being calibrated. While in January a 3 week window (cal 5) provides a more robust climatology although fair RPSS similar to the 2 week window (cal 3), in April a 3 week window climatology (cal 5) seems to introduce noise which results in worse fair RPSS.

These results show that employing a window to increase sample size can lead to a more robust climatology estimate and a better calibration, however depending on the time of the year there is a limit on the size of the window, after which skill scores are degraded. The optimal length is found to be 3 start dates corresponding to 2 week window.

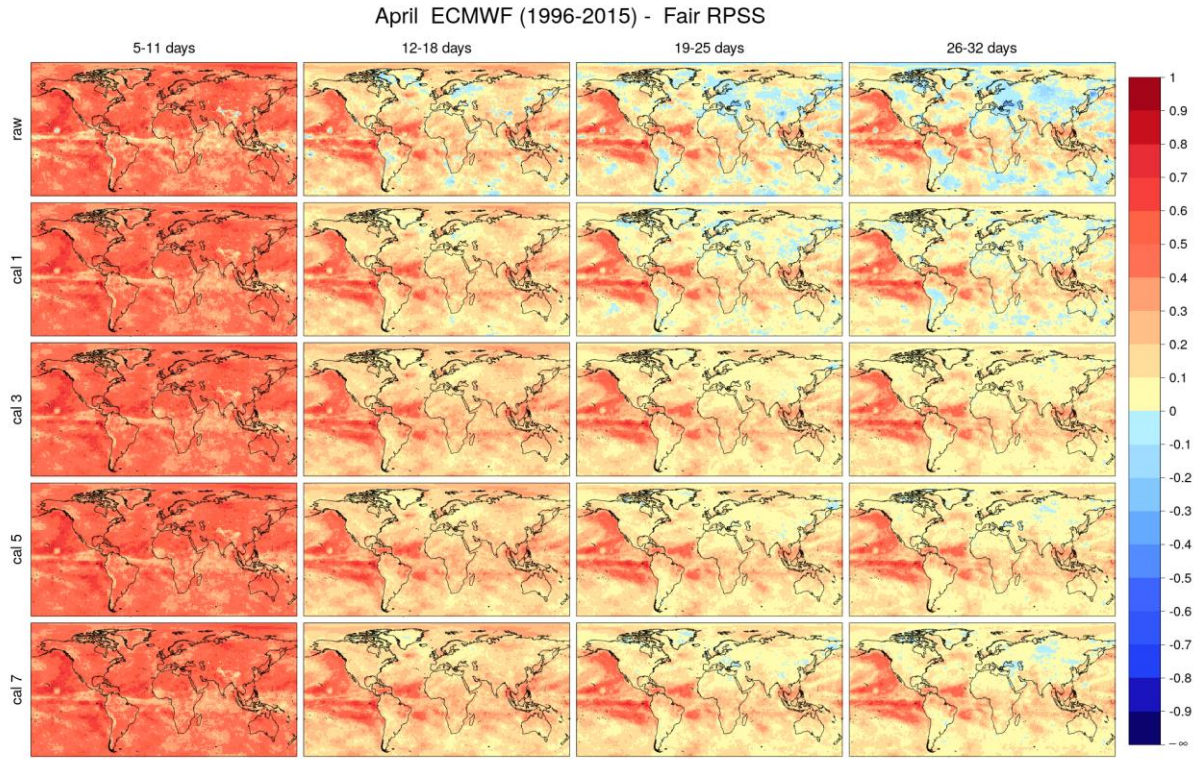


Figure 6. Similar with Figure 5 but for April.

4.4. Conclusions

To produce skillful probabilistic sub-seasonal predictions from a system like NCEP-CFSv2 it is advisable to produce a lagged ensemble by pooling together forecasts initialised at different times. To better understand the influence of the different choices of the lagged ensemble in probabilistic skill, several tests were performed for 2m temperature. It is concluded that 3-day lagged ensemble is a good compromise between the benefit of increased number of ensemble

members and the degradation of including older forecasts. A 3-day lagged ensemble has been used by other studies considering NCEP CFSv2 (Wang & Robertson, 2019).

The effect of the variance inflation calibration method is very dependent on the size of the hindcast. By increasing the NCEP-CFSv2 hindcast from 12 years to 20, skill is enhanced especially in weeks 3 and 4. This is achieved by merging to the 1999-2010 hindcasts to the forecasts issued from 2011.

Finally, the use of a running window to increase the hindcast sample size for the computation of the climatology for operational implementation of the variance inflation calibration on ECMWF-MFS has been tested, simulating an operational context. The optimal window for calibration has been found to be 2 weeks, obtained by employing 3 start dates. This approach is therefore recommended to be applied operationally in the DST.

5. Improving skill – Recommended Enhanced bias adjustment

5.1. Introduction

In this section we focus on the analysis of bias adjustment methodologies to improve the skill scores of the forecast. Indeed, both weather and climate forecasts, not to mention climate projections, are affected by biases resulting from the imperfect ability to numerically reproduce the processes that are responsible for climate variability (e.g. Torralba et al., 2017, Zhao et al., 2017).

The analysis is conducted on the seasonal forecast and for two daily variables: temperature and precipitation. Those variables are essential per se but also central to renewable energy evaluation. In order to bring a recommendation for an operational implementation in the DST, we evaluate:

- the skill of the reference methods (including the one used in the S2S4E DST) applied on monthly and daily forecast values,
- the skill of a new method based on quantile mapping applied on the same data.

The data we use in this work is, as in other deliverables, the 25 member seasonal hindcast of the ECMWF System 5 forecast model from January 1993 to December 2016. These are daily values at 1° resolution as delivered by the C3S Climate Data Store. As for observations we use the ERA5 reanalysis interpolated on the same spatial and temporal resolution. To benchmark the skill of the methods we used the CRPSS and RPSS on terciles as probabilistic skills compared with climatology, and the ACC of the ensemble mean as deterministic skill.

As for the criteria to evaluate the methods we decided to adopt a user centered approach by selecting the following conditions:

- perform the benchmark on the daily adjusted values
- focus on lead month 1 and onwards since lead month 0 is covered by the sub seasonal forecast,
- prefer the method that shows significantly higher positive skill (>5%).

We thus perform the analysis on the first three lead times that are included in the DST: Lead 0 (or Month 0) as a benchmark, and Leads 1 and 2 (or months 1 and 2).

5.2. The reference methods

We benchmark the two reference methods presented in Torralba et al. (2017) and also present in the benchmark made by Manzanas et al. (2019). The methods are referred to as “Simple bias adjustment” and “Calibration”. In the previous studies the authors focused on 10m wind speed for the winter quarter (Torralba et al., 2017) and on precipitation and temperature for boreal winter and summer quarters (Manzanas et al., 2019). In the present version of the DST, the Calibration method is used for both the sub-seasonal (applied weekly) and seasonal and forecast (applied monthly). Here we applied the methods on monthly but also daily data in order to explore skill for daily adjusted forecast for the first time in the S2S4E project. To do so, the coefficients of the equations are estimated as monthly averages or monthly statistics and

applied on daily forecast values (see equations in Annex 2). We used the CDO libraries and coded the equations to apply on the forecast data.

When the reference methods are applied for precipitation negative values appear at daily, but also monthly averages, for some dry days or months around the globe. This problem was also reported for monthly wind surface values with the Calibration method in a study by Lledó et al. (2019). As mentioned in the later study, setting the negative precipitation values to 0 alters the monthly means and thus affects the bias adjustment. However, since having positive precipitation values is an prerequisite and improving the reference methods for daily values being out of scope of this work, we decided to simply set to zero the negative values despite possible effects on the monthly bias but probably minor on the skill scores.

5.3. The CDFt method

Quantile mapping (QM) is rapidly becoming the method of choice by operational agencies to post-process raw GCM forecast (Zhao et al., 2017). QM techniques have the advantage to be more suitable as they are less prone to introduce negative values for non-continuous variables as precipitation. However, while QM methods are highly effective in correcting bias, they have also been criticized for not ensuring reliability or guarantee coherence (Zhao et al 2017).

The Cumulative Distribution Function-transform (CFDt) method (Michelangeli at al., 2009, Vrac et al., 2012, Vrac et al., 2016) is a non-parametric quantile mapping-based technique that was never tested on seasonal forecast. It was developed for adjusting (or downscaling) climate change projections. It corrects model values in a future period given observations and model data in a reference (or historical) period. CDFt has the advantage of accounting for climate change (or changes in the underlying distribution) by estimating through a mathematical transformation the future adjusted CDF to infer a future quantile map instead of using the quantile map of the historical period to adjust the future as in simple QM methods. CDFt has been extensively applied to adjust daily data but also 3 hourly data of climate change projections (Bartok et al 2019). It is worth noting that a specific processing is applied to precipitation in order to coherently adjust both intensities and occurrences of precipitation (Vrac et al., 2016). When applying the CDFt method at the ensemble, we choose to mix all the ensemble members (Crochemore et al., 2016; Ilias Pechlivanidis, personal communication). We thus apply a QM specific for each month and each lead time for all daily values from every member in this month. We note that in Zhao et al. (2017), the QM matching takes place at the level of individual ensemble members and for monthly values.

5.4. Results

For Temperature, the uncorrected forecast shows positive annual CRPSS (Figure 7) and RPSS (Figure 8) as well as significant ACC on lead 0 (Table 3) that decreases rapidly in the following leads. Skill is present mostly on oceanic regions with a maximum in the tropics while continents tend to show lower skill. CRPSS becomes negative over continents while RPSS remains globally positive for all leads. The visual comparison of annual skill maps shows some small differences between methods in favor of the CDFt adjusted forecast over the tropical oceans. The domain spatial averages (see Table 3) however confirm that differences among methods are not significant (i.e. less than 5% points).

The same analysis with the methods applied monthly shows that differences are significantly better for the monthly CDFt adjusted forecast while not significant for the reference methods (Figure A 7, Figure A 8 and Table A 1 in Annex 2). Overall, there is a decrease in skill for the CDFt method when applied on daily values compared to monthly making the skill similar for all methods. In terms of ACC there are no significant differences nor between methods, nor between monthly and daily adjustment (not shown).

For precipitation, the uncorrected forecast shows highly negative annual CRPSS (Figure 9) and relatively low ACC (below 0.5, see Table 4) on lead 0 that decreases rapidly in the following leads. RPSS is positive only on Lead 0 becoming strongly negative afterwards (Figure 10). In this case, methods try to correct a forecast with very poor skill. All methods fail to recover significant positive skill except over a small tropical oceanic belt. The visual comparison of annual skill maps shows that the Calibration method is the one that achieves the highest skill improvement but without reaching positive skill. Differences with the other methods are higher and thus significant for CRPSS and less so for RPSS. The domain's spatial averages (see Table 4) confirm the quantitative differences among methods: the calibration method improves skill the most without reaching positive skill. Again, these results of the reference methods applied on precipitation should be taken with care considering the negative values issue mentioned previously (negative values in the reference methods are set to zero before calculating the skill scores).

The same analysis with the methods applied monthly shows that positive skill, although modest (below 10%), is reached for Lead 0 (Table A 2 in Annex 2). There are no significant differences (5% and above) except over Africa for the CDFt method. Skill in the following leads remains negative as for the daily adjusted forecast although with smaller values.

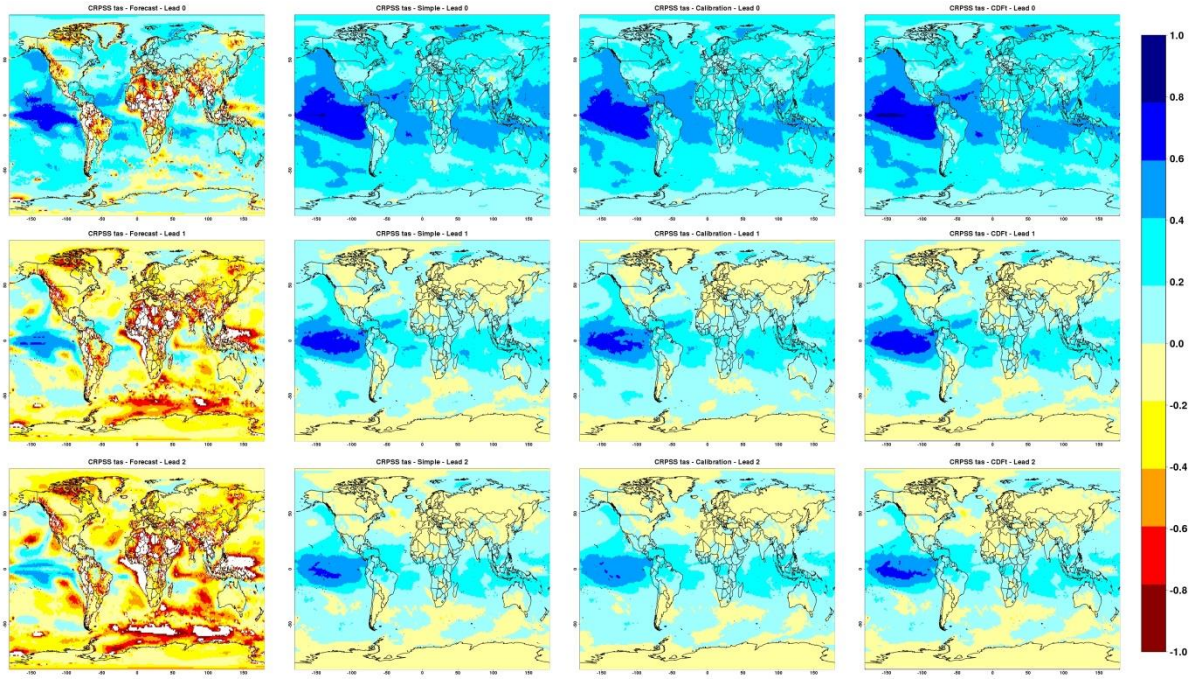


Figure 7. **Annual CRPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the raw forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).**

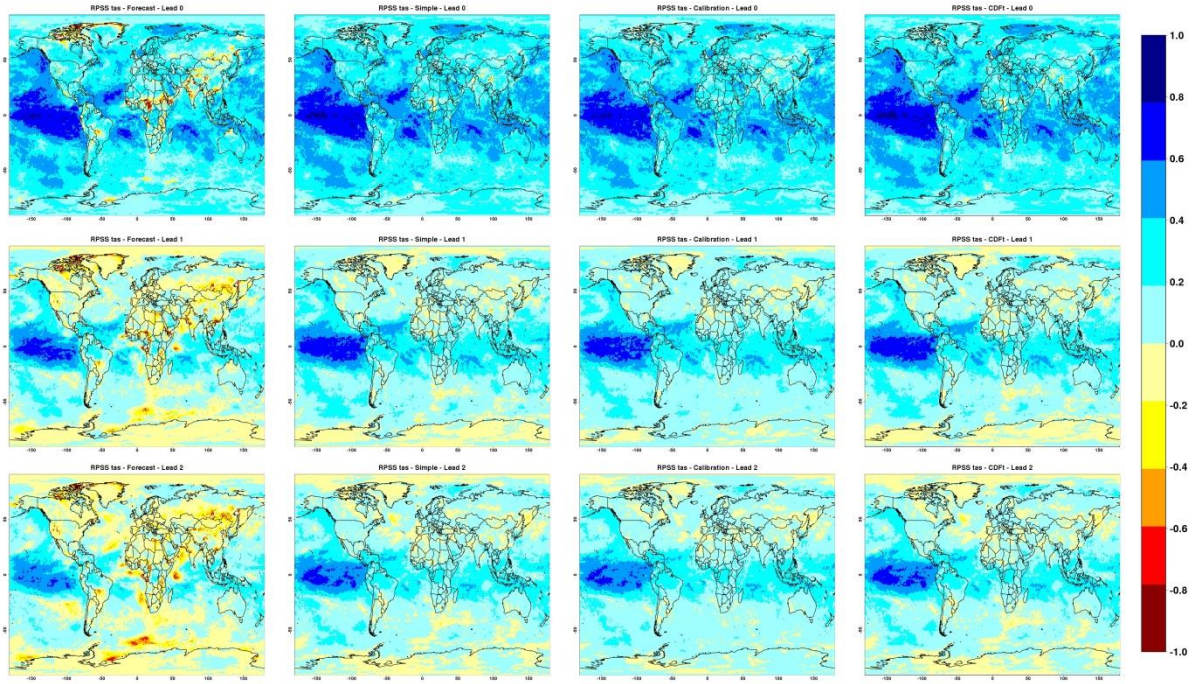


Figure 8. **Annual RPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).**

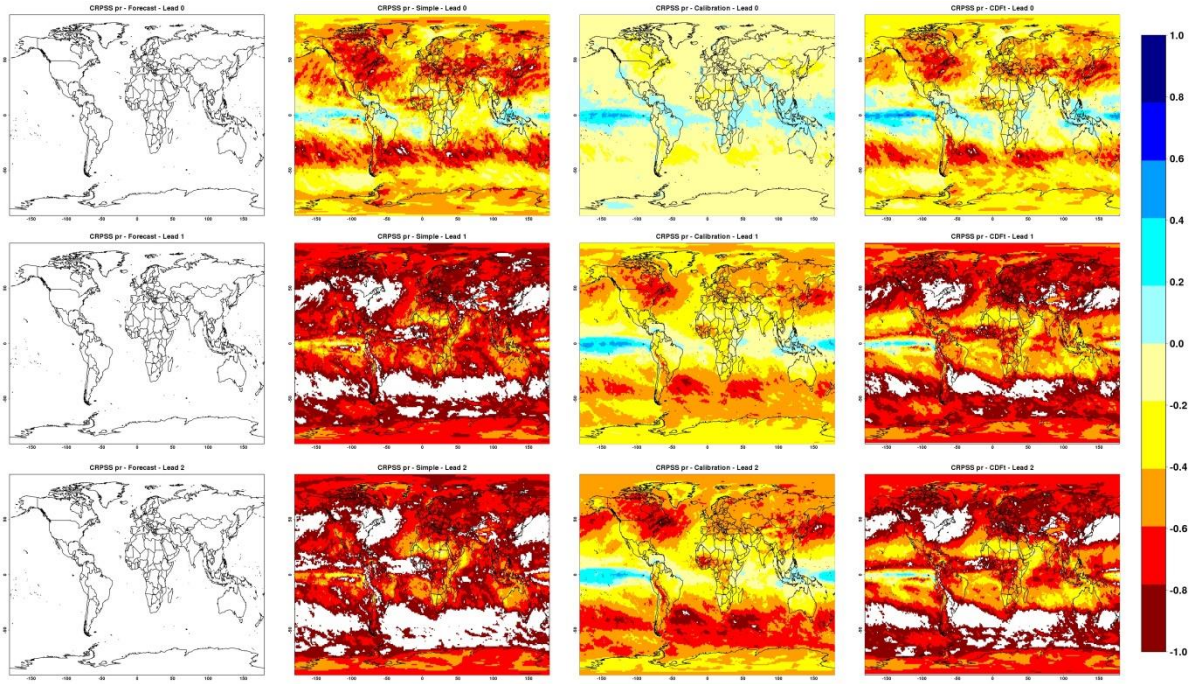


Figure 9. **Annual CRPSS score of daily averages for precipitation for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).**

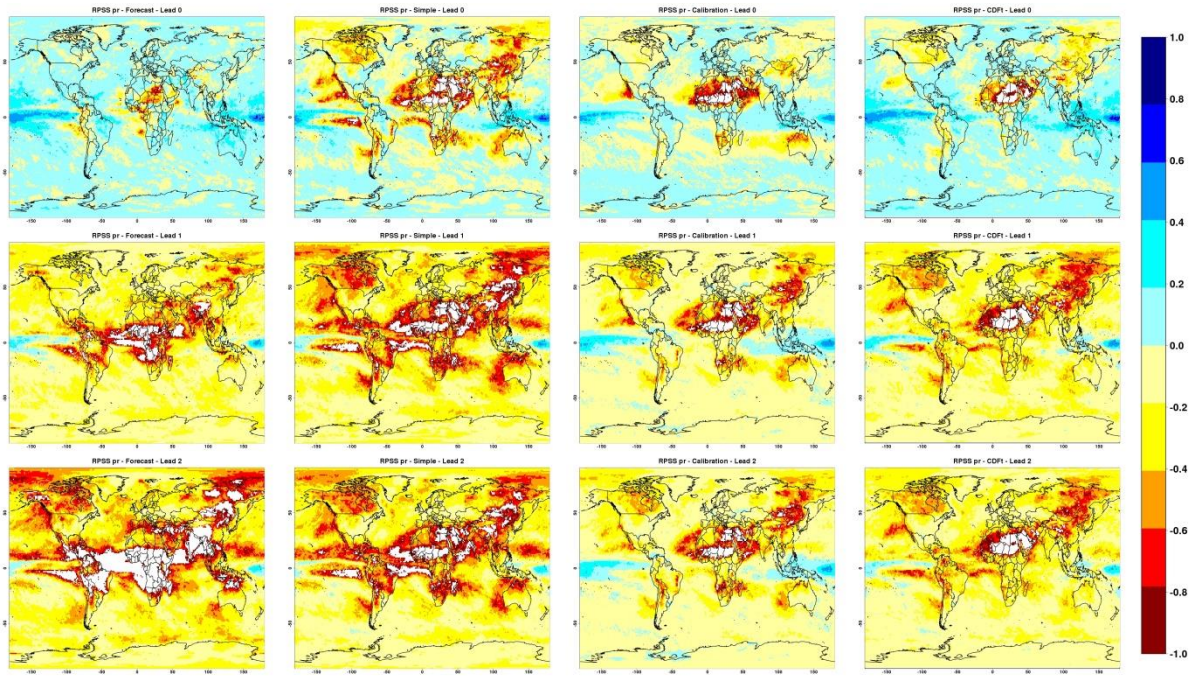


Figure 10. **Annual RPSS score of daily averages for precipitation for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted daily with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).**

		Lead 0			Lead 1			Lead 2		
Domain	Forecast	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**
Europe	Forecast	-0.01	0.67	0.22	-0.27	0.27	0.00	-0.32	0.21	-0.04
	Simple	0.23	0.68	0.26	-0.01	0.27	0.05	-0.03	0.22	0.03
	Calibration	0.22	0.68	0.25	0.01	0.29	0.06	-0.01	0.23	0.04
	CDFt	0.23	0.68	0.26	-0.01	0.27	0.05	-0.02	0.22	0.03
Africa	Forecast	-0.49	0.73	0.17	-0.67	0.47	-0.01	-0.73	0.40	-0.06
	Simple	0.28	0.73	0.32	0.09	0.47	0.13	0.05	0.41	0.10
	Calibration	0.28	0.74	0.32	0.10	0.48	0.14	0.07	0.42	0.11
	CDFt	0.29	0.74	0.32	0.09	0.47	0.14	0.05	0.41	0.10
East Asia	Forecast	-0.17	0.75	0.21	-0.46	0.42	0.01	-0.53	0.37	-0.04
	Simple	0.30	0.75	0.30	0.07	0.43	0.11	0.04	0.37	0.08
	Calibration	0.29	0.75	0.29	0.08	0.44	0.12	0.06	0.38	0.10
	CDFt	0.30	0.75	0.31	0.07	0.43	0.11	0.04	0.37	0.08
North America	Forecast	0.04	0.73	0.26	-0.25	0.43	0.03	-0.33	0.37	-0.01
	Simple	0.31	0.74	0.34	0.08	0.43	0.11	0.04	0.37	0.09
	Calibration	0.30	0.74	0.33	0.09	0.44	0.12	0.06	0.38	0.10
	CDFt	0.31	0.74	0.34	0.08	0.43	0.12	0.05	0.37	0.09

Table 3. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for 2m temperature daily averages for the first three lead times of the uncorrected forecast and the forecast adjusted daily with the three different methods spatially averaged over Europe, Africa, East-Asia and North America.

		Lead 0			Lead 1			Lead 2		
Domain	Forecast	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**
Europe	Forecast	-51.57	0.46	0.04	-165	0.06	-0.22	-285.3	0.03	-0.36
	Simple	-0.45	0.46	-0.08	-0.79	0.05	-0.31	-0.77	0.03	-0.29
	Calibration	-0.07	0.42	0.00	-0.37	0.06	-0.13	-0.46	0.03	-0.15
	CDFt	-0.37	0.45	0.02	-0.75	0.06	-0.22	-0.79	0.03	-0.21
Africa	Forecast	-59.38	0.47	-0.06	-183.0	0.14	-0.59	-309.1	0.11	-1.11
	Simple	-0.34	0.48	-0.48	-0.68	0.16	-0.70	-0.68	0.13	-0.68

	Calibration	-0.07	0.45	-0.31	-0.28	0.18	-0.48	-0.34	0.15	-0.47
	CDFt	-0.21	0.47	-0.21	-0.52	0.15	-0.50	-0.55	0.13	-0.52
East Asia	Forecast	-54.17	0.49	0.06	-174.8	0.15	-0.34	-300.5	0.13	-0.67
	Simple	-0.44	0.49	-0.17	-0.85	0.16	-0.49	-0.87	0.14	-0.48
	Calibration	-0.07	0.44	-0.05	-0.32	0.16	-0.22	-0.39	0.15	-0.24
	CDFt	-0.32	0.48	-0.01	-0.71	0.15	-0.30	-0.77	0.13	-0.31
North America	Forecast	-50.21	0.44	0.04	-164.7	0.09	-0.25	-288	0.08	-0.46
	Simple	-0.55	0.44	-0.20	-0.89	0.09	-0.44	-0.90	0.09	-0.42
	Calibration	-0.12	0.39	-0.08	-0.43	0.10	-0.23	-0.51	0.10	-0.24
	CDFt	-0.41	0.42	-0.04	-0.76	0.09	-0.29	-0.81	0.09	-0.29

Table 4. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for precipitation averages for the first three lead times of the uncorrected forecast and the forecast adjusted daily with the three different methods spatially averaged over Europe, Africa, East-Asia and North America.

5.5. Discussion and Conclusions

Overall, the results depend on the initial skill of the uncorrected forecast. They are thus related to the forecasted variable (temperature has more skill than precipitation) but also to the temporal resolution of the adjustments (monthly adjustment has slightly more skill than daily adjustments). When skill is present, as for the monthly adjusted temperature, all three methods improve the skill. In this case the CDFt method seems to perform systematically better for Lead 0, although the difference is non-significative. For the Lead 1 and 2 skill decreases and differences between methods is non-significative. On the other hand, when there is no or very low skill in the uncorrected forecast, the methods can improve it but without reaching positive values, as in the daily adjustment of precipitation. In this case, the Calibration method showed a significant reduction of negative skill compared to the other methods.

Considering the criteria adopted in the introduction (focusing on positive skill in Lead 1 and above) no method is significantly better in terms of skill scores. This is in phase with the benchmark done when adjusting the forecast quarterly (Manzanas et al., 2019). However, it should be noted that the reference methods introduce negative values of precipitation when applied monthly or daily while the CDFt method with specific preprocessing doesn't (keeping in mind that the CDFt method is more resource intensive). This property of the CDFt method is the only reason to favor it for precipitation adjustment as eventually for any other non-continuous variable (e.g. wind and solar).

6. Improving skill – Towards a multi-model approach

6.1. Introduction

The current skill of S2S forecasting in Europe is limited due to a low inherent predictability and limited quality of models and observations (Arnal et al., 2017). This limited predictability arises from a combination of errors and uncertainty from meteorological S2S forecasts, initial hydrological conditions used to initialize the forecast, impact model structure and parametrization and downscaling errors. Forecasts from a single climate system could provide limited information in a specific region, however statistical post processing tools can be used to improve skill and further to estimate the uncertainty affecting the forecasted variable from multiple-model systems. When applying statistical post-processing methods to a multi-model system the aim is to take advantage of the individual model strengths in different climatic regions and seasons, e.g. by assigning different model weights to the models based on past forecasting performance.

In the case of hydrological applications, the two main components of a seasonal forecasting system are the atmospheric model and the hydrological (impact) model, with the former providing the input for the latter. Improved simulations of forecast uncertainty in a hydrological forecast system can be obtained by: 1) using perturbations in the initial conditions, 2) using different model structures, 3) using parametric uncertainty, or 4) making use of different meteorological forcing datasets. There is currently an increasing demand for multi-model ensemble systems that have multiple meteorological and impact models, to achieve improved accuracy and consistency in the forecasting operations. In seasonal meteorological forecasting, there has seen an increase in the ensemble sizes of multi-model forecasts, whereas hydrological forecast systems are still provided with single impact model setups.

A number of methods have been developed and proposed in order to post-process and combine forecasts from multiple climate and/or hydrological models (Wanders et al., 2019; Klein et al., 2016). The methods allow estimation of the predictive uncertainty and aim to increase forecast reliability and sharpness. Among the many investigations, Muhammad et al. (2018) assessed various multi-mode approaches using statistical post-processing techniques to improve the predictability of seasonal streamflow forecasts in the Canadian Prairie region. Similarly, Wanders and Wood (2016) applied different methods to post-process seasonal temperature and precipitation forecasts from multiple seasonal forecasting systems. One of the methods that has been applied in various fields is the Bayesian Model Averaging (BMA; Raftery et al., 2005). BMA is used to estimate multiple model predictions and their uncertainty through a weighted mean of the predictive distributions of individual models.

In this chapter, we assess various statistical post-processing methods which allow combining S2S forecasts from different climate models and for different variables of interest related to the energy sector. The results and insights are specific to the investigated variable, whilst the methodology followed is not consistent through all the investigations, since individual in-house expert knowledge drove the investigation.

6.2. Solar radiation and temperature

6.2.1. Introduction

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 analyzed a selection of prediction systems (Table 5) from the Copernicus C3S seasonal forecasts product (C3S dataset; https://climate.copernicus.eu/seasonal-forecasts).

Center/Prediction system	Atmosphere/ land model	Ocean/sea ice model	Initialization
ECMWF SEAS5	IFS Cycle 43r1 grid: TCO319 L91	NEMO v3.4 grid: ORCA025 L75	atmosphere/land: ERA-Interim/ECMWF oper. analysis ocean: ORAS5/ORTA5
MF SYS6	ARPEGE v6.2 grid: TL359 L91	NEMO v3.6 grid: ORCA1 L75	atmosphere/land: ERA-Interim/ECMWF oper. analysis ocean: MERCATOR-OCEAN
DWD SYS2	ECHAM 6.3.04 grid: T127 L95	MPIOM 1.6.3 grid: TP04 L40	Atmosphere/land: ERA-Interim/ECMWF oper. analysis ocean: ORAS5/ORAS5

Table 5. Model configuration, resolution, and initialization strategy of each contributing system.

One-month lead retrospective seasonal predictions are collected for the considered models for the period 1993-2014 (1st May and 1st November start dates, i.e. June-July-August, JJA and December-January-February, DJF); the validation period has been chosen according to the availability of satellite observations of GLCF-GLASS surface albedo 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 (2m temperature T2M, surface solar radiation downward SSRD, moisture convergence -Qdiv). We analyzed 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 most of the applications of seasonal forecasts are based. A new metric is 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). The methodology

developed is applied to compare a subset of the seasonal prediction systems included in the C3S multi-model ensemble. This analysis provides information on how to optimize the selection and combination of the models for the most relevant target variables for energy applications.

6.2.2. Process-based model inter-comparison

The deterministic skill (anomaly correlations, ACC) of the three selected models from C3S 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 Meteo France predictions tend to be better over East US and West Europe (Figure 11a). 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 Meteo France. Over that region there is a known problem in the ECMWF system related to the ocean initialization with ORAS5 (Johnson et al., 2019). The affected region is centered on a box defined by the longitudes 50-30W and the latitudes 45-55N 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 analyzed 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 [hereinafter normalized yearly covariance, $r(x,y)_i$], 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 11b). Positive (negative) values of normalized yearly covariance differences means 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.

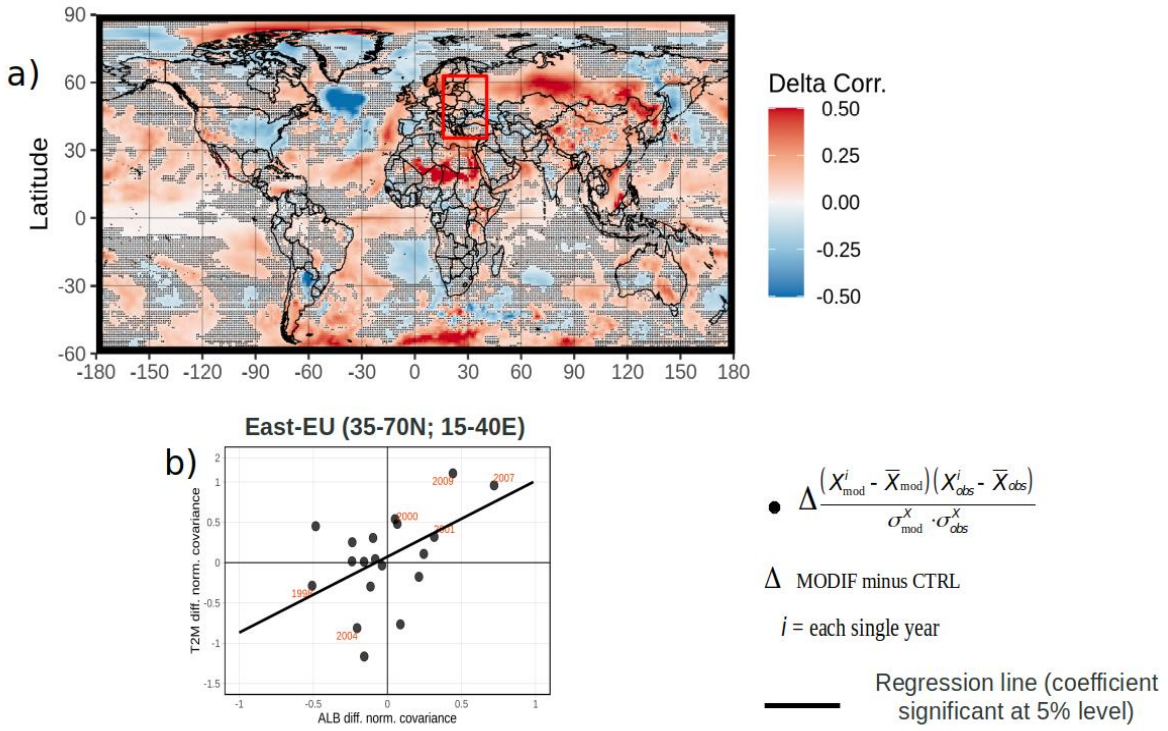


Figure 11. a) 1-month-lead boreal winter (DJF) 2m temperature ECMWF minus Meteo France correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and Meteo France for the predictions averaged 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 same analysis has been applied to compare ECMWF and DWD systems (Figure 12a). 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 12b) 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 Meteo France for 1-month lead seasonal hindcasts for boreal summer (June-July-August, JJA) surface solar radiation (Figure 13a) shows that the ECMWF model is performing better over North America, East Europe and Central Asia while Meteo 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 13b).

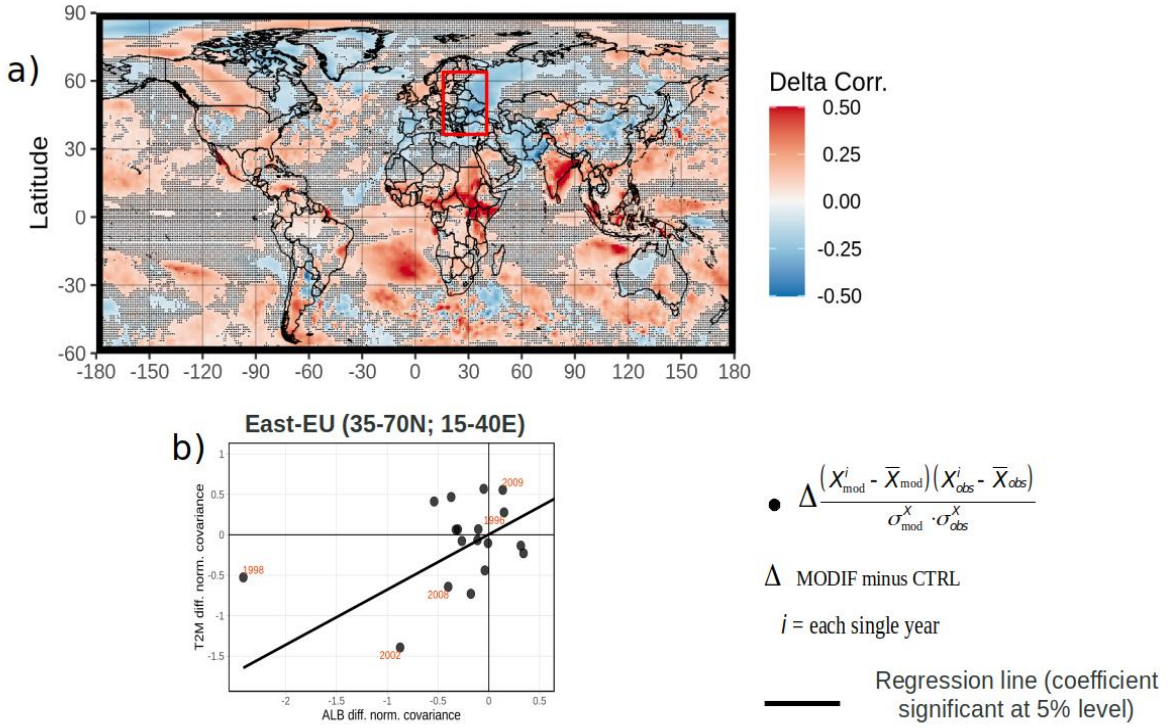


Figure 12. a) 1-month-lead boreal winter (DJF) 2m temperature ECMWF minus DWD correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and DWD for the predictions averaged over the East-European domain (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 14a. 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 14b) 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.

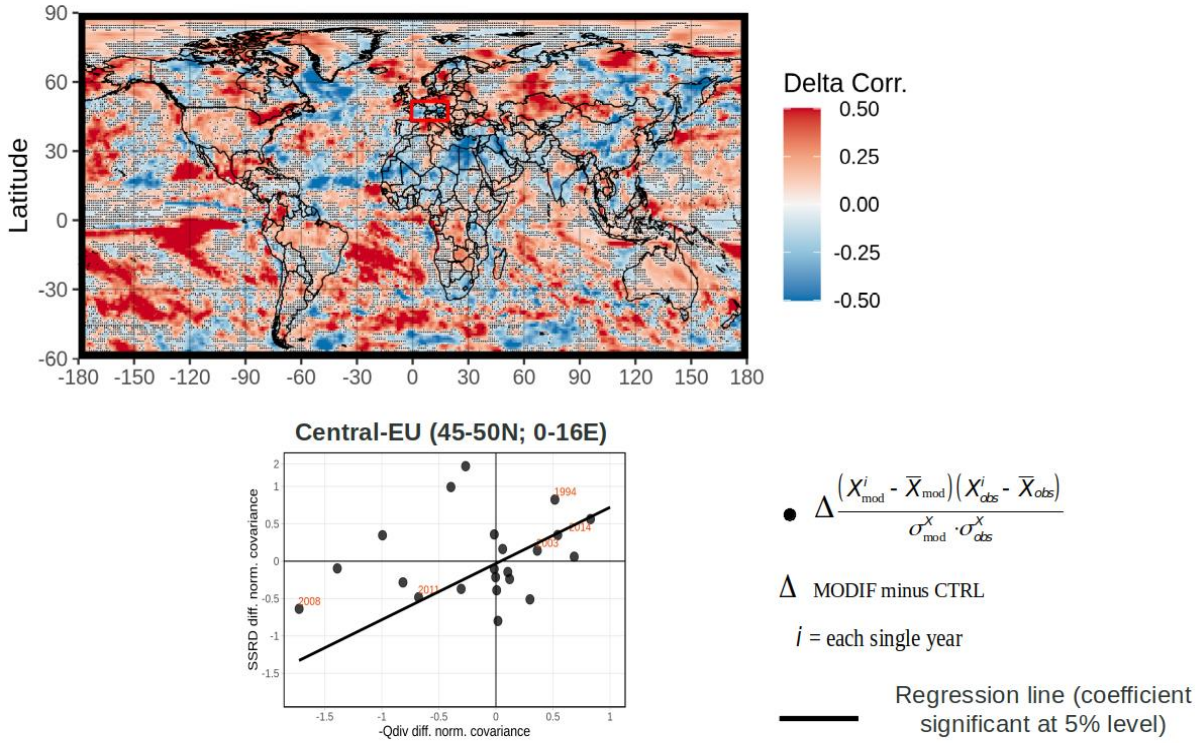


Figure 13. a) 1-month-lead boreal summer (JJA) surface solar radiation downward ECMWF minus Meteo France correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and Meteo France for the predictions averaged over the Central-European domain (0E–16E; 45N–50N) 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).

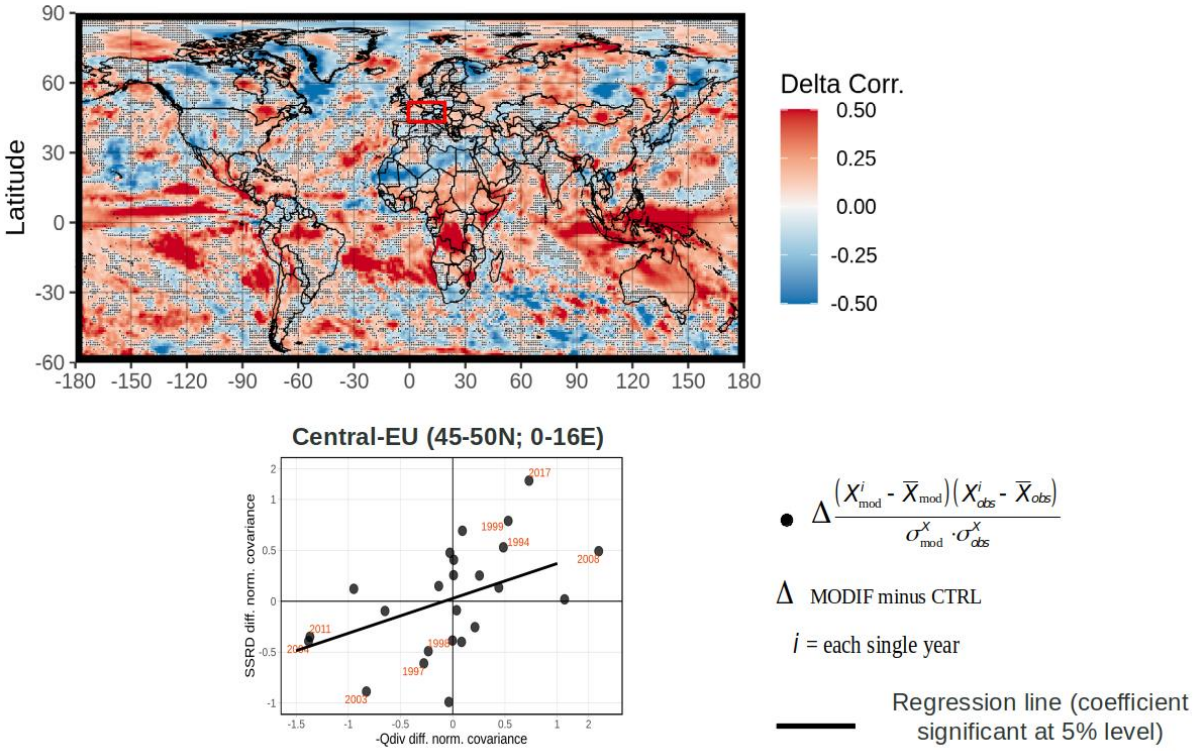


Figure 14. a) 1-month-lead boreal summer (JJA) surface solar radiation downward ECMWF minus DWD correlation difference vs. ERA5. Dotted grid points did not pass significance test at 10 % level. b) Scatterplot of the normalized yearly covariance differences between ECMWF and DWD for the predictions averaged over the Central-European domain (0E–16E; 45N–50N) 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).

6.2.3. Probabilistic scores and model independence

The probabilistic accuracy has been analyzed in terms of Brier Skill score (BSS) for dichotomous events of conditions being above/below upper/lower terciles of the sample distribution. Furthermore, starting from the definition of the Brier score (Equation 1; Wilks, 2009), we have developed a new metric, the Brier score covariance (BScov; Equation 2), which estimates the relative independence of prediction systems 1 and 2:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2 \quad (1)$$

$$BS_{cov} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i^1 - o_i)(y_i^2 - o_i)}{\sqrt{BS^1 \cdot BS^2}} \quad (2)$$

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. Superscripts (^{1}) and (^{2}) in Equation (2) 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. BScov 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, BScov takes into account 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 which is the more relevant for the project. 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 Meteo France are concentrated over East Europe and in general at the higher latitudes while Meteo France system is performing better over the Iberian peninsula and the Mediterranean countries (Figure 15a). 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 15b). The BScov metric has been used to assess the relative independence of the selected models in the probabilistic information they provide (Figure 15c and d). For both combinations, the larger probabilistic independence (lower BScov values) is over the ocean, indicating that model or initialization differences in this component play a major role that must be taken into account 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).

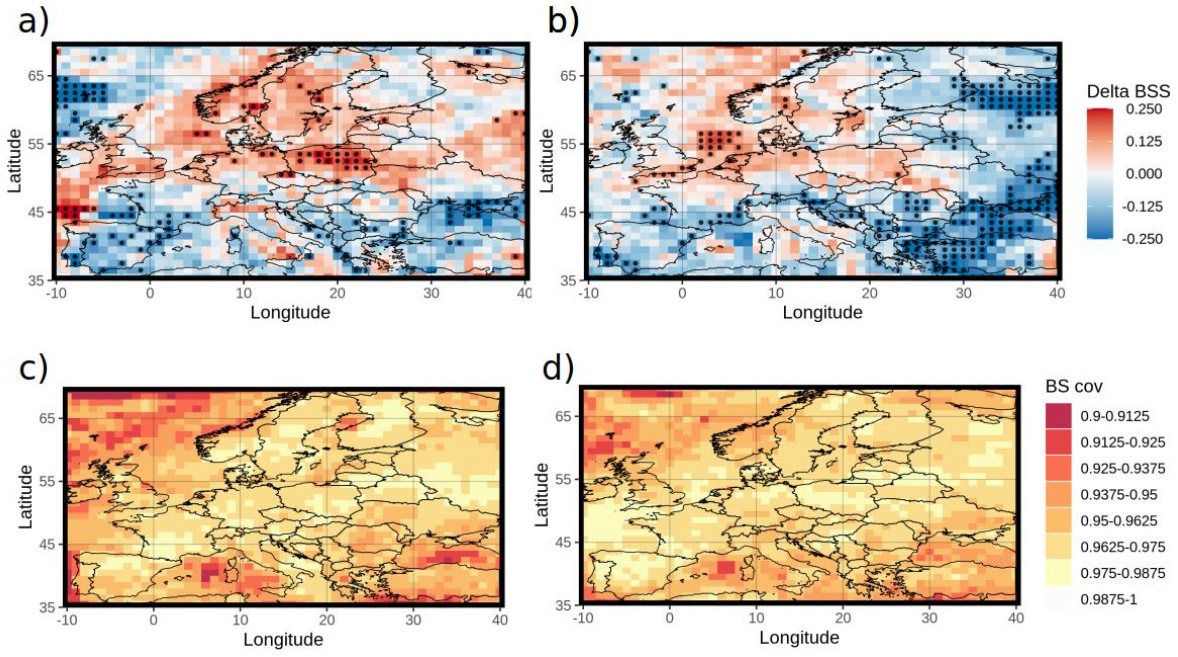


Figure 15. Spatial distribution of the BSS differences of the probabilistic forecasts for below-normal (below lower tercile of sample distribution) 2m temperature in Boreal winter (DJF) for Europe domain. (a) ECMWF minus Meteo France; (b) ECMWF minus DWD;

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 Meteo France; (d) ECMWF vs DWD.

Figure 16 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 Meteo France over land at the higher latitudes while the latter model is outperforming over Central Europe (Figure 16a), consistently with the deterministic analysis in previous section. Comparison of ECMWF vs DWD (Figure 16b) 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 16c and d), again we see a large contribution of the ocean component. The fact that ECMWF and Meteo 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.

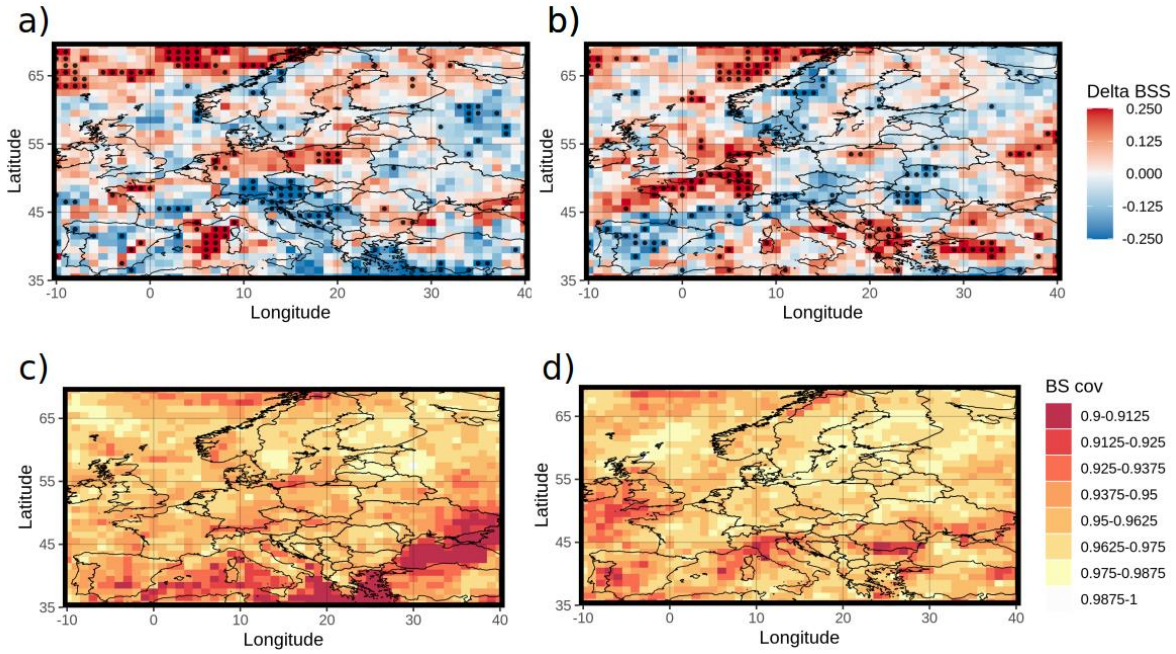


Figure 16. Spatial distribution of the BSS differences of the probabilistic forecasts for below-normal (below lower tercile of sample distribution) surface solar radiation downward in Boreal summer (JJA) for Europe domain. (a) ECMWF minus Meteo France; (b) ECMWF minus DWD; dotted are the areas that passed a significance test at the 10% level. Probabilistic independence as measured by the new BScov metric: (c) ECMWF vs Meteo France; (d) ECMWF vs DWD.

6.2.4. Discussion and Conclusions

A selection of three models (ECMWF, Meteo France and DWD) from the Copernicus C3S dataset has been analyzed in terms of both deterministic and probabilistic skill for the start dates of May (boreal summer season) and November (boreal winter season). Two target variables of interest for the energy users have been considered: 2m temperature and surface solar radiation downward. Skill differences between the prediction systems have been analyzed together with related possible sources of predictability. Attribution analysis evidenced the importance of snow-albedo processes for temperature predictions in DJF and the effect of the atmospheric dynamics through moisture convergence for the prediction of surface solar radiation in JJA. A new metric, the Brier Score Covariance, has been developed to quantify the probabilistic independence among the models aiming at optimizing model selection and combination strategies for the MME. Given that the potential benefit that can be gained by using a MME approach increases with the independence of the contributing systems, the application of the BScov metric enhances the potential to increase the performance of the MME combinations. In particular, the relative degree of independence of the prediction systems, together with their skill, provide key information on which models might be taken into account in the MME for the development of a climate service for a specific region and a specific target variable. The number and selection of models that perform better is usually different depending on the region and the energy sector under consideration. A combination strategy currently under investigation considers two-steps: (i) rank/filter the models for a specific variable over the region of interest based on the skill; (ii) combine the models selected in step (i) based on weights defined by model independence.

6.3. A hydrological investigation

6.3.1. Methodology

For the hydrological investigation, three seasonal climate forecasting systems from the institutions Météo France (MF), UK MetOffice (GLOSEA5), and ECMWF (SEAS5) are considered in the multi-model analysis. All these systems use the same reference meteorological dataset (HydroGFD; Berg et al., 2018) for bias-adjustment, whilst the bias-adjusted products are forcing the E-HYPE hydrological model (Hundecha et al., 2016) to calculate river streamflow (m^3/s). The modified version of the Distribution Based Scaling (DBS; Yang et al., 2010) bias-adjustment approach is used to account for drifting conditioning the bias adjustment on the lead time. The bias-adjustment is done for each month with the estimated parameters being applied to each one of them. All years within the period of interest are used to estimate the DBS parameters. All the above choices are made in order to include different state-of-the-art seasonal prediction systems of different spatiotemporal reliability patterns and a state-of the-art bias-adjustment method in the multi-model analysis. An overview of the seasonal prediction systems and their characteristics is given in Table 1.

Here we used the three dynamical seasonal prediction systems for a period of 21 years (1994–2014) over the pan-European domain to form our multi-model ensemble seasonal meteorological forecasting system. The multi-model ensemble is based on the equal model averaging (EMA) approach, where all three single systems are equally weighted, independently of the number of ensemble members each one has. All systems produce seasonal atmospheric forecasts at the beginning of each month for the upcoming 7 months with a daily temporal

resolution. The historical atmospheric (precipitation and temperature) forcing that is based on the HydroGFD dataset (Berg et al., 2018) is used as a reference (perfect forecast) in the multi-model ensemble. The initial E-HYPE hydrological conditions used in the forecasting framework are obtained from the historical E-HYPE simulation based on the HydroGFD meteorological input.

The performance of the river streamflow forecasts at the basin scale is evaluated based on the fair version of the Continuous Ranked Probability Skill Score (CRPSS). Fair CRPSS compensates for the effect of the number of members on the score because it rewards ensembles with members that behave as if they and the verifying reference are sampled from the same distribution. Here, we use the simulated climatology as benchmark which is the historical simulated climatology of the variable of interest. A number of historical time-series simulated from the targeted forecast date are randomly selected from the model climatology (1994-2014) to build an ensemble of possible outcomes. Daily streamflow is averaged over monthly periods and used to assess the model forecasting skill. In terms of the temporal scale analysis, we assess the skill for lead months 0-5 (lead month 0 is the month of the initialization, so for a forecast launched in November, lead month 0 is November).

6.3.2. Results

We analyzed the monthly streamflow at the basin scale over the entire European domain in order to understand the emerging spatiotemporal patterns and limits of predictability. According to Pechlivanidis et al. (Under Review in Water Resources Research) the patterns of predictability in streamflow are related to both the memory of the river systems and the climatic properties.

Figure 17 and Figure 18 present the CRPSS score for lead month 0 and for the winter and summer months respectively, whilst the figures include results for all systems (MF, GloSEA5 and SEAS5) and the multi-model (EMA). As expected, the skill varies both geographically and seasonally with acceptable skill in different regions for lead month 0. The seasonal streamflow forecasts using the MF climate model are very skillful in norther Europe in the winter months; however, the forecasts are not skillful in central Europe. In winter, central Europe shows good skill levels under the GloSEA5 model. Although the forecasting skill is not very high (i.e. > 0.7), the forecasts generated using the SEAS5 model seem to perform adequately at most river basins in Europe. The EMA analysis is generally taking advantage of the individual model strength and generates forecasts that perform the better than any individual contributor in terms of CRPSS values and spatial coverage. The EMA forecasts are skillful over most European domain.

The skill of the streamflow forecasts based on the MF and GloSEA5 systems is higher in the summer than in the winter months. In addition to this, more regions have positive skill in summer than winter. It is interesting that MF-based streamflow forecasts in central Europe that lacked skill in winter are skillful in the summer. For the GloSEA5-based streamflow forecasts, the summer skill is extended to northern Europe and also the Mediterranean. The regions that were skillful in winter under the SEAS5 model are still skillful in summer. The skill for this system only seems to increase in the summer period. Finally, the skill levels of the multi-model again outperform the individual systems both in terms of skill levels and its spatial coverage.

Moreover, Figure 19 shows the CRPSS of the ensemble streamflow forecasts of the individual models together with the post-processed EMA forecasts for the lead months 0-5 for all

European sub-basins. The distribution of the CRPSS values for all the sub-basins are illustrated as boxplots. Results show that despite the high forecasting skill in lead month 0, the skill deteriorates rapidly with increasing lead month. In addition, as shown also in the maps above, the EMA multi-model method can outperform all individual models and in all lead months. Even when the individual models show negative skill (as mean value), the forecasting skill from the EMA method is positive, i.e. lead month 3 in autumn, lead month 2 in winter, lead month 1 in spring and lead month 2 in summer. It is also very interesting to note that the spatial spread in forecasting skill (described as the spread in the boxplot) is more condensed for the EMA results, highlighting that the method can even introduce skill at regions that would lack forecasting skill. Finally, results show that summer and autumn have higher skill than in winter and spring in all the systems (and that could indicate a higher predictability). That also applies to the spatial spread of the results in skill. A similar analysis using all available systems for every single month can be found in Figure A 9 in Annex 3.

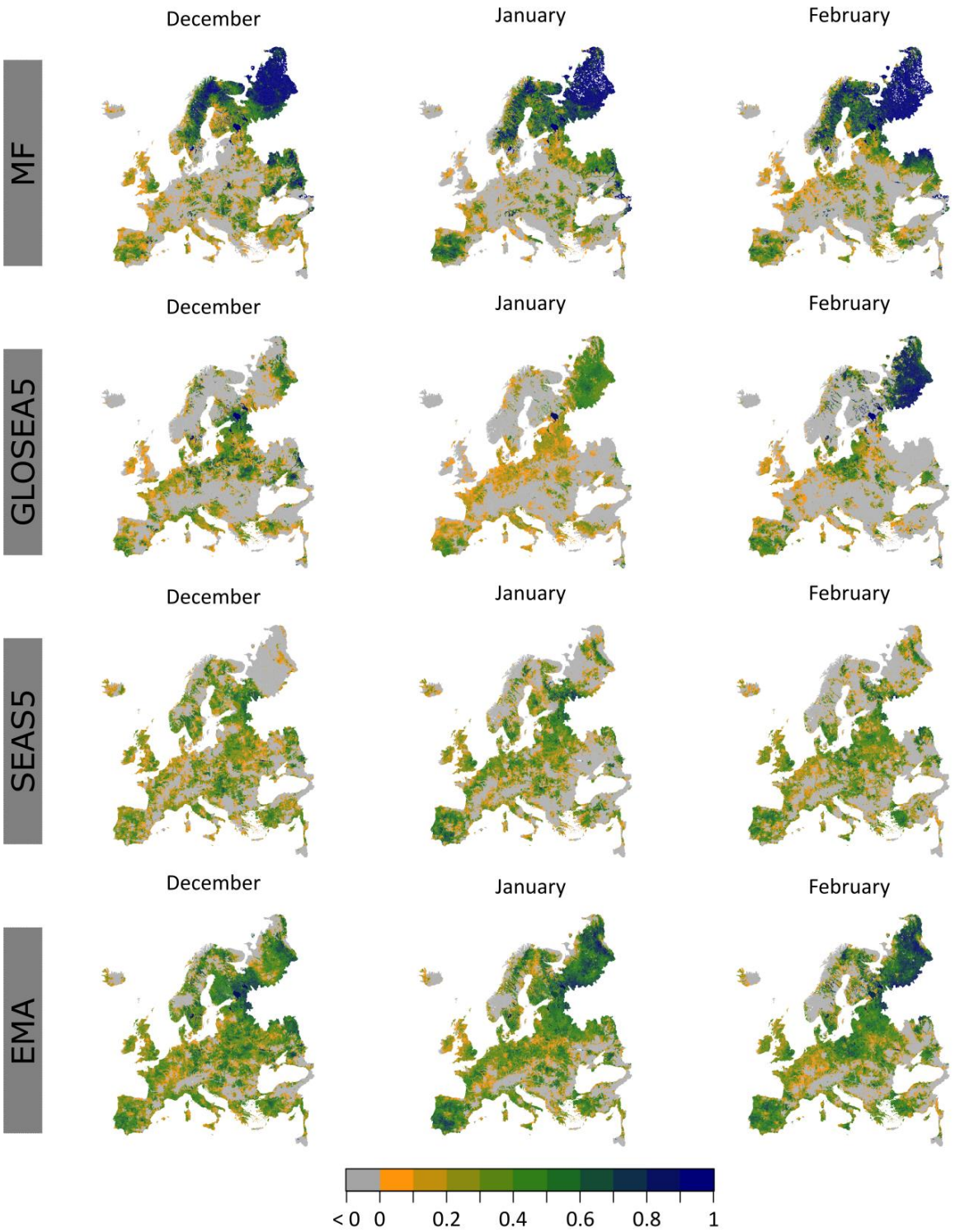


Figure 17. Lead month 0 monthly spatial variability of the CRPSS of streamflow for the winter months.

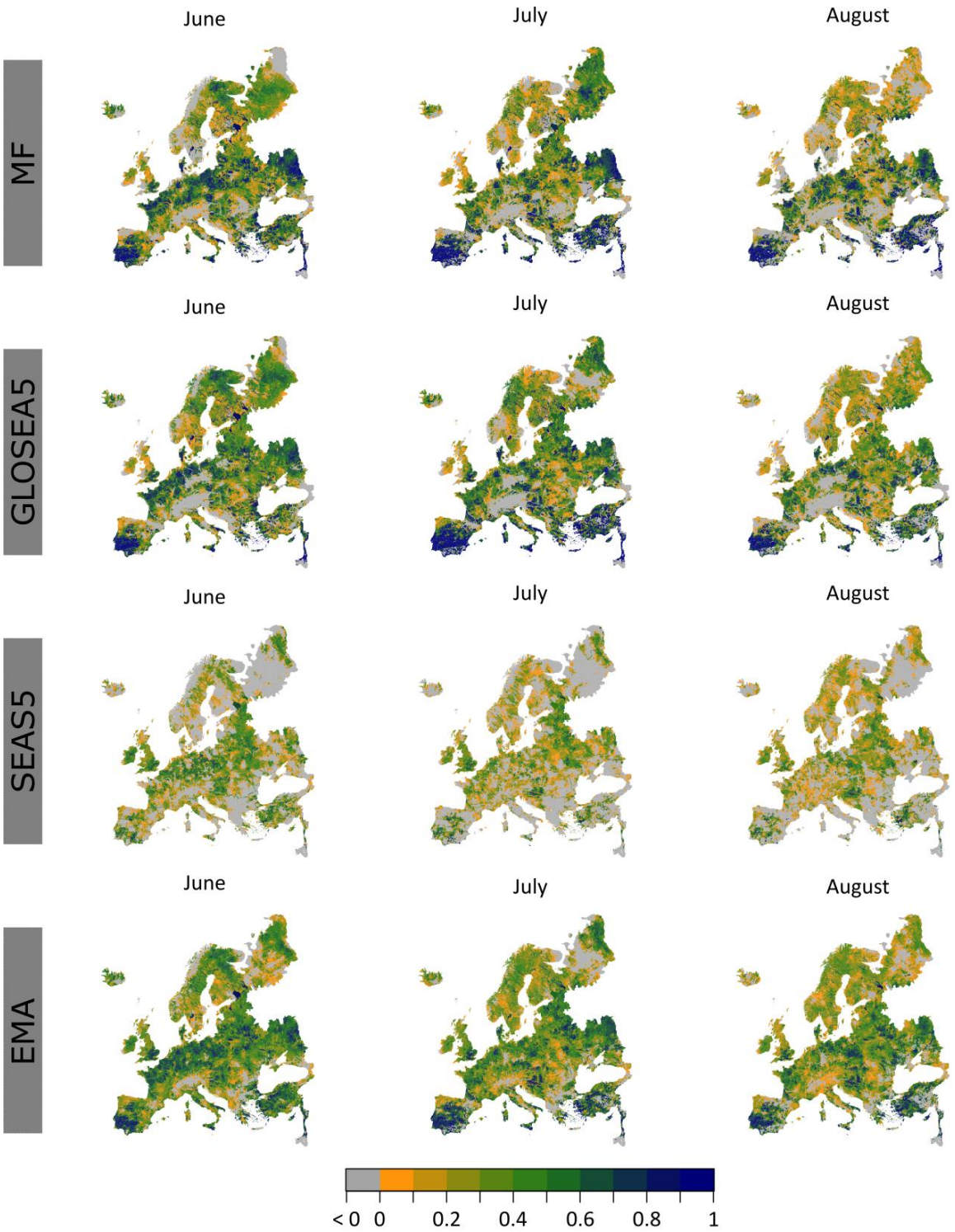


Figure 18. Lead month 0 monthly spatial variability of the CRPSS of streamflow for the summer months.

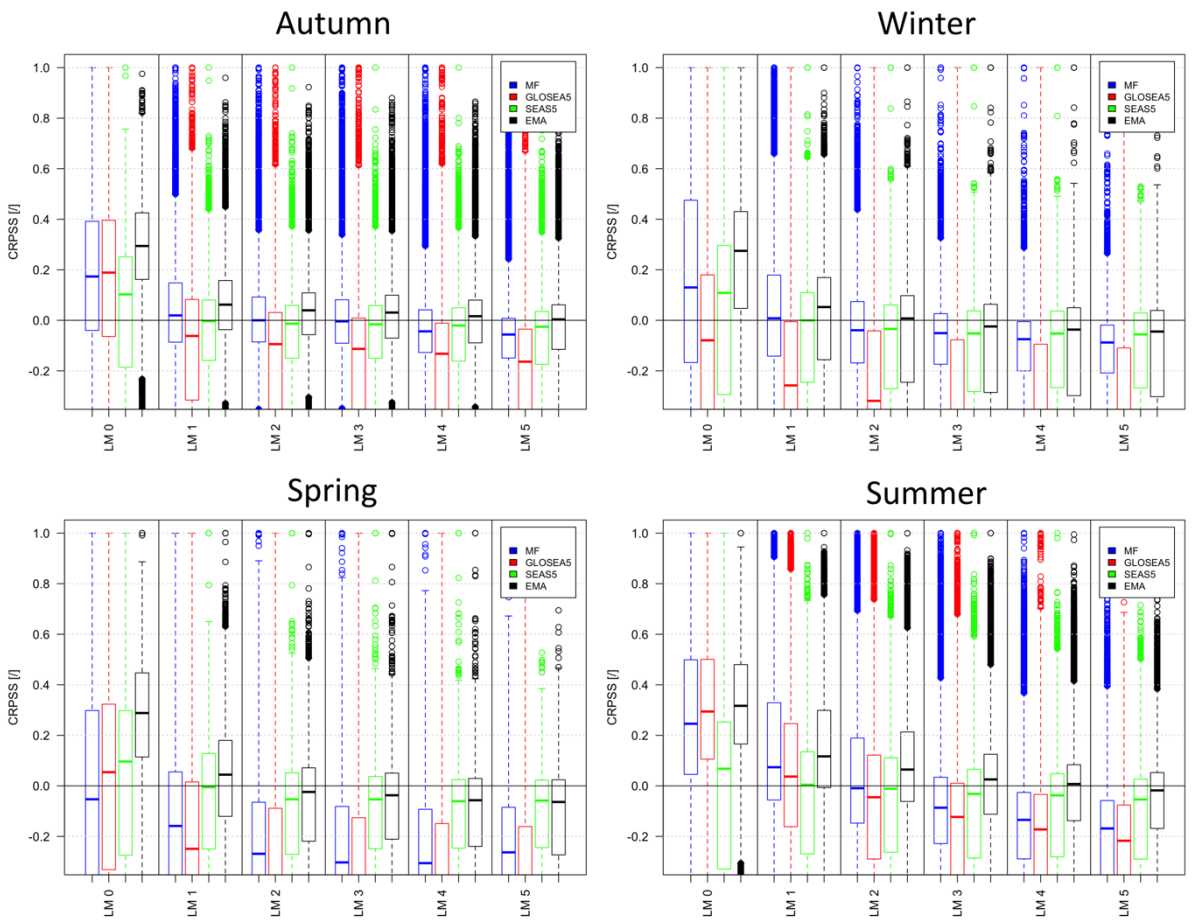


Figure 19. CRPSS values of individual systems (MF, GLOSEA5 and SEAS5) and the multi-model (EMA) seasonal streamflow forecasts for all European sub-basins. The boxplots and outliers are based on results from all 35408 sub-basins in the E-HYPE model setup.

6.3.3. Discussion

The quality of seasonal streamflow forecasts relies on a forecasting chain that includes at least seasonal meteorological forcing (singular or multiple systems), initialisation of hydrological model states and a hydrologic model setup (or a set of setups). To improve the forecast quality and further the decision-making, this chain can be advanced by introducing additional components that allow assimilation of data to set the initial model states (e.g. in-situ / earth observations of soil moisture, snow water equivalent), post-processing of seasonal meteorological forecasts (e.g. bias-adjustment), and post-processing of hydrologic forecasts (e.g. conditioning to local data). Currently forecast service development is ad hoc with improvements made to single parts of the forecasting chain when and where available, and with only very limited guidance on the relative importance of each component to the forecasting chain performance. For example, the investigation conducted here, only accounted for the seasonal atmospheric forcing without addressing other sources of predictability.

To date only few investigations identified the dominant sources of predictability in seasonal hydrological forecasting (i.e. the initial hydrological conditions and meteorological forcing) at the continental and global scale; however these only consider different forcing data, model setups and benchmarking (Greuell et al., 2019; Li et al., 2009; Shukla and Lettenmaier, 2011;

Yossef et al., 2017, 2013; Zhang et al., 2017). The lack of large-sample studies across a variety of modelling settings at multiple spatiotemporal scales and under changing environmental conditions has limited the understanding of how skill evolves in space and time. Sensitivity analysis methods have been used at place-based investigations for seasonal streamflow forecasting to: 1) identify critical lead times after which the streamflow skill mainly depends on the meteorological forecasts (and less on initial conditions) (Wood and Lettenmaier, 2008), and 2) quantify the increase in hydrological skill as a results of increasing the skill in one of the dominant sources (Arnal et al., 2017; Wood et al., 2016). Application of these methods at the continental/global scale can better exploit our understanding of the sources of predictability in seasonal predictions supplying the users/stakeholders with evidence to guide forecast developments.

Finally the hydrological model (setup, data, structure, parameters) is another source of uncertainty. Here the analysis was conducted using pseudo-observations as reference, which are not always comparable to real observations, but provide complete information in the spatial and temporal domain. Nevertheless, the assessment against pseudo-observations reduced model errors from the analysis to the minimum, and hence results targeted the hydrological processes despite the actual model performance (Bierkens and Van Beek, 2009; Van Dijk et al., 2013).

6.3.4. Conclusions

Here, we show that regardless of the seasonal prediction system used as forcing in a hydrological model, the quality of the seasonal streamflow forecasts varies both geographically and seasonally, whilst the skill depends on the season and deteriorates with increased lead months. The highest skill levels over Europe overall are shown in summer and autumn in comparison to winter and spring, and these results are consistent for all systems. More importantly, this investigation highlights that the multi-model approach, even if this was based on a simple equal model weighting method, can improve the seasonal streamflow forecasting skill consistently. Even when the individual systems do not show skill, the multi-model method is capable of achieving positive skill values; this was observed in high lead times. In addition, the multi-model approach can improve the spatial variability of the skill over the entire domain; more regions achieve positive skill.

6.4. Energy country average

6.4.1. Introduction and Motivations

The benefits of multi-model combinations in climate forecasting have been previously introduced and described for different temporal scales (e.g., Siebert and Stephenson, 2019, DelSole, 2007, Sansom et al., 2013). Most typical combination methodologies involve weighting strategies that assign each model a constant factor, either uniformly or through a skill assessment. Given that the skill of the models can vary at different timescales, and for multiple reasons (for example, seasonally varying skill, or due to changes in the forecasting system), the fact that these weights remain constant introduces limitations.

Within the realm of Machine Learning, a family of algorithms has been developed to perform 'online prediction with expert advice' (Cesa-Bianchi et al., 2006). These methods consider a set of 'experts' which, after being initially weighted uniformly, produce subsequent predictions in which the combination rule or 'mixture' is updated to optimize a loss or skill function.

These online prediction methods have several potential advantages for their use in climate prediction:

- The fact that the expert combination is updated in every forecast step allows the system to adjust in certain conditions (e.g., the ones mentioned above) to preserve skill;
- Since a different combination can be easily obtained for different quantiles of the predictand distribution, a robust system can be trained that maximizes skill for its full range.
- The risk of including incompetent or counterproductive experts is minimized by the fact that the mixture is able to adapt and discard them (or assign them minimal weights).

Another potential application of these online prediction methods could be on the design of 'seamless' forecasting systems in the sub-seasonal to seasonal sense, which is of interest to this project. For example, the system could be trained with a set of experts that include subsequent launches of a sub-seasonal forecast as well as prior launches of a seasonal forecast. If at any point there is useful information arising from the longer lead time seasonal forecast, the mixture would assign higher weights to it.

A set of these online prediction methods has been tested here, and compared to more typical multi-model combination techniques to assess their usefulness for the prediction of country-level energy demand, and potentially other variables. The following sections will show that these innovative methods exhibit significant skill improvements (higher than 5%) with respect to more standard techniques and to individual forecasting systems for lead weeks up to 4.

6.4.2. Methodology

This study considers country-level energy demand time series from two s2s numerical weather prediction (NWP) systems (ECMWF ENS-ER and a lagged version of NCEP CFSv2) as well as the ERA5 reanalysis (Hersbach et al., 2019) for the common period 1999-2010. Demand series were obtained using a model (Bloomfield et al., 2019) that is forced by 2 metre temperature and only accounts for the weather-driven variability and disregards any socio-economic drivers, which typically reflect as cycles (e.g. weekly) and trends. The hindcasts from the two NWP systems were combined and their skill to represent the demand corresponding to ERA5 was evaluated for averages over weeks 1 through 4, according to the project convention.

A simple set of experts was considered, split into two categories: NWP-based and reanalysis-based. The NWP-based set included:

- Quantiles of the ensemble distribution (Q10, Q35, Q50, Q65, Q90): The quantiles of the hindcasts ensembles calculated for each start time and for each system: ECMWF (_1) and lagged NCEP (_2).
- Minimum and Maximum of the ensemble distribution (FCST_MN, FCST_MX): For each start, the min and max values from both hindcast systems are used, with the potential to retain the seasonality of the hindcast systems.

The reanalysis-based set includes:

- Quantiles of the climatology (Q10_CLIM,..., Q90_CLIM): ERA5 climatology for each hindcast day calculated using a leave-one-out approach on the years. The daily time series were then smoothed using a 15-day normal kernel filter.
- Persistence (PERS): Weekly persistence forecast based on ERA5, calculated using days -7 to 0, being 0 the hindcast start date.

- Last-year persistence (PERS_1yr): Weekly persistence forecast based on ERA5's demand for the same week of the prior year.
- Seasonal minimum and maximum (SEAS_MN, SEAS_MX): For each hindcast week, the min and max values in the ERA5 climatology are obtained using a leave-one-out approach on the hindcast year.

These sets of experts were combined, fully or partially, using two 'mixtures' or aggregation methodologies: Bernstein Online Aggregation (BOA; Wintenberger 2017) and Polynomial Potential Aggregation (MLpol; Gaillard, Erven, and Stoltz, 2014). The methodologies were implemented using the Open Source R Opera Package (https://cran.r-project.org/web/packages/opera/opera.pdf). Each mixture was applied using all experts (BOA, MLpol) and only the NWP-based experts (BOA_NWP and MLpol_NWP) in order to assess the presence of any added value from including the reanalysis-based predictors. Each aggregation rule was applied to represent the hindcast distribution using a set of quantiles ('Qgrid'). In this study, we considered a quantile spacing of 0.05 (5%). This means that the resulting predictions constitute an ensemble of 19 members, evenly separated in quantile space from 0.05 to 0.95. Additionally, a different aggregation rule was created for each hindcast lead (weeks 1 to 4).

The multi-model mixtures were then benchmarked against a set of standard references:

- Climatology (CLIM): The climatology forecast was built for each quantile in 'Qgrid' using the 11years of ERA5 climatology with a leave-one-out approach.
- Uniform mixture (UNIF_NWP): For each quantile in 'Qgrid', a quantile forecast is obtained from each system and then they are averaged with equal weight. Note that this reference forecast is not directly using the expert set, unless for the specific quantiles listed above.
- Oracles: These oracles are the optimal combinations of all the NWP-based experts in the set through a multiple linear regression approach, under two constraints: in the convex case (O_NWP_conv), the weights range between 0 and 1 and have to add up to 1. In the linear case (O_NWP_lin), the weights range freely. Oracles require the full knowledge of all the evaluation period and therefore provide an upper boundary for the skill of a performance-based linear multi-model combination. Most typically, these combinations would be trained using an out-of-sample period.

The skill of the obtained mixtures was also compared to that of the individual NWP systems.

6.4.3. Results

We introduce here the performance of the online aggregation methods for country-level United Kingdom weekly demand. The initial period 1999-2010 (12 years) gets reduced to 2000-2010 since the first year is lost to generate 1-year persistence forecasts. Initial tests revealed that it takes around 2 years for the aggregation methodologies to achieve a quasi-equilibrium in their weights variability. Therefore, all skill evaluations in this section are restricted to the 9-year period 2002-2010.

We first present the evaluation of the multi-model methodologies for Q50 (i.e., their deterministic skill). Figure 20 presents the average pinball losses (also known as quantile losses) corresponding to Q50, which by definition coincides with the mean absolute error (MAE) of the predictions. Panel a) shows that MLpol and BOA, respectively, are the most skillful combinations for predicting week 3. They are followed closely by the oracle combinations, the BOA_NWP and MLpol_NWP mixtures and UNIF_NWP and CLIM references.

Figure 20b presents the losses of a subset of models relative to the UNIF_NWP combination. It shows that for week 3, MLpol and BOA show approximately a 5% increase in skill. All the model combinations and the climatology present an increase in skill with respect to UNIF_NWP for every lead time. Regarding the individual NWPs, ECMWF is more skillful than the benchmark only for week 1 and NCEP is always worse. The figure also shows that the mixtures that exclude the experts from the reanalysis (BOA_NWP, MLpol_NWP) have higher losses associated to them.

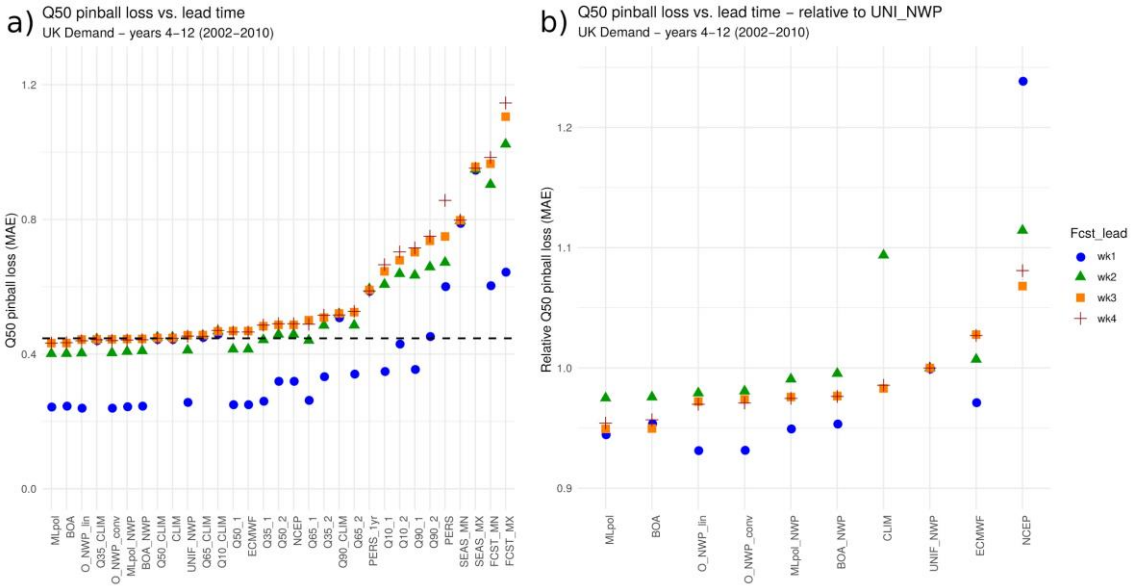


Figure 20. a) UK demand Q50 average pinball loss associated to the aggregation rules, the individual experts and the reference forecasts. The average losses for weeks 1 to 4 are presented as different symbols. The items on the x-axis are sorted from smaller to bigger loss based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for climatology, as a reference. b) Pinball losses for a subset of the models and references, but expressed as relative losses with respect to the UNIF_NWP combination, for each corresponding week.

To explore the composition of the aggregation methods, Figure 21 presents the time-average weights for each lead. It shows that for week 1, the higher weights are assigned to the quantiles of the ECMWF model. As lead time increases, the mixtures that include reanalysis-based experts shift weight to the quantiles of the climatology, whereas the NWP-based mixtures assign larger weights to the quantiles of NCEP than in shorter leads.

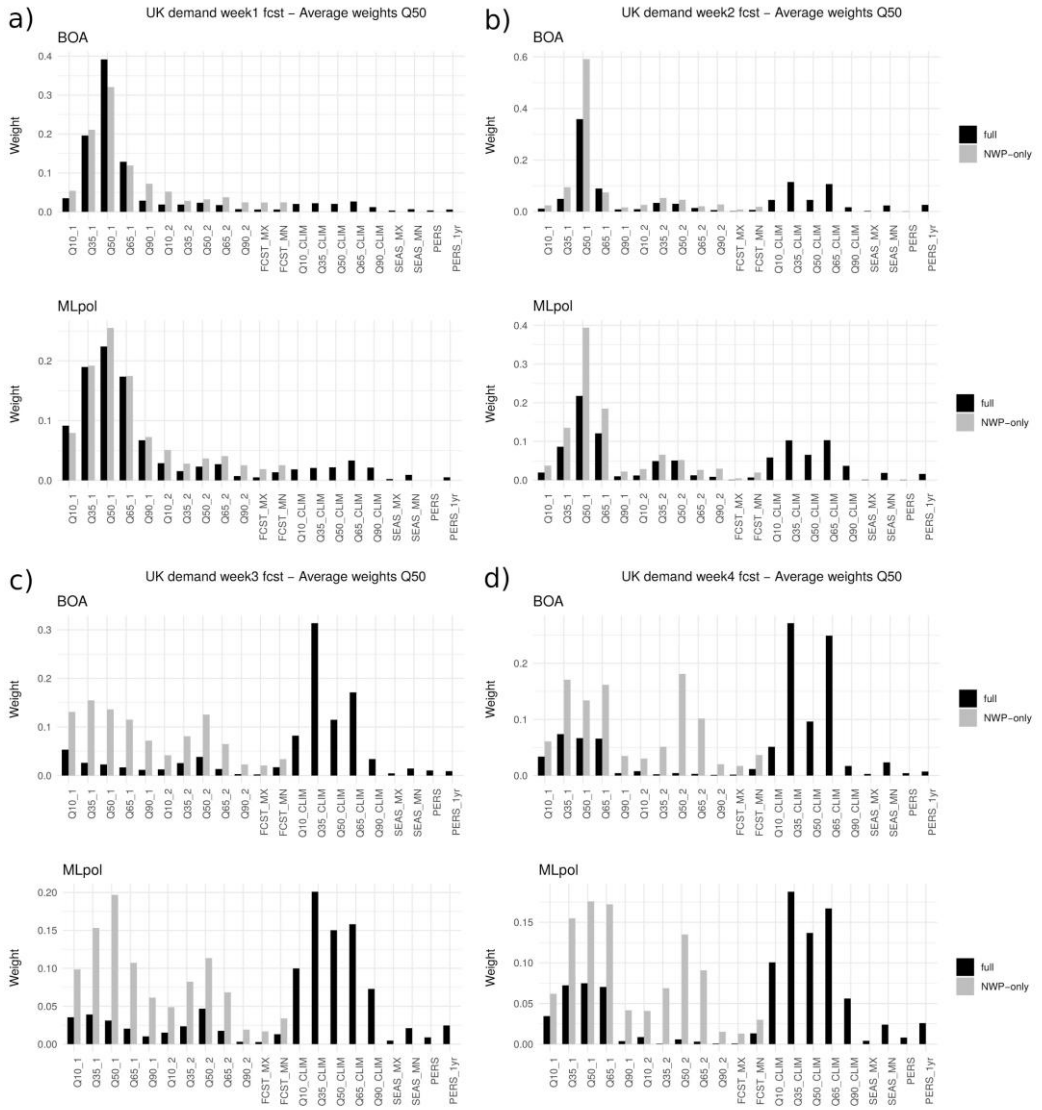


Figure 21. Time-average weights involved in the Q50 aggregations: BOA (black bars, top panels), MLpol (black bars, bottom panels), BOA_NWP (grey bars, top panels) and MLpol_NWP (grey bars, bottom panels); for leads: a) week1, b) week2, c) week3 and d) week4.

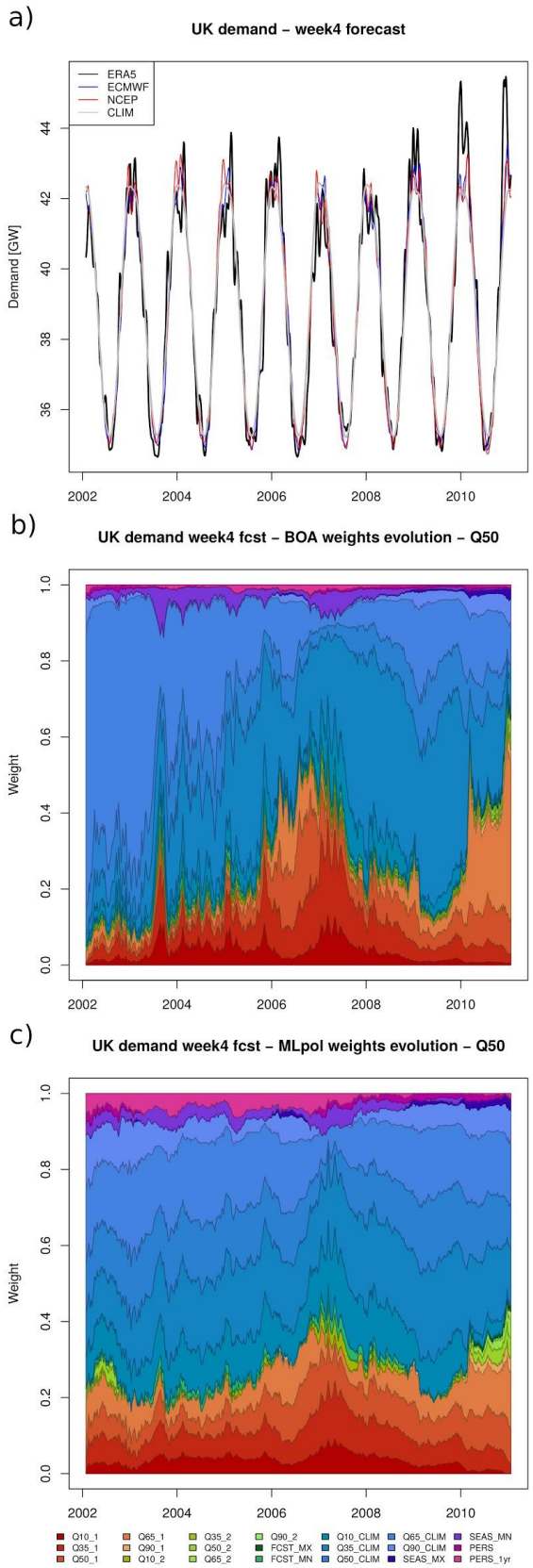


Figure 22. a) Temporal evolution of UK demand for week 4. The black line corresponds to reanalysis, the grey line corresponds to the leave-one-out climatology and the red and blue lines are the ensemble means for ECMWF and NCEP, respectively. Weight evolutions for the aggregation rules corresponding to that same lead time for a) BOA and b) MLpol.

As in fact the mixtures are characterized by their time-varying weights, it is interesting to explore their evolutions. Figure 22 presents the temporal evolution of UK demand and the BOA and MLpol weights for the week 4 hindcasts. One of the outstanding features of the weight evolutions is an apparent 'shock' centered around 2007. A close inspection reveals that the aggregations (panels b and c), which were assigning the largest weights to climatology quantiles by that week, drop these weights in favor of ECMWF quantiles. It is possible to see that around that time, the reanalysis showed lower demand that the climatology (panel c), anomalously warm fall and winter) and therefore the climatology was not as skillful as a forecast. The BOA and MLpol methods were able to 'learn' this fact early in the season and adjusted the weights accordingly. Because the actual demand was lower than the ECMWF ensemble mean, the largest weights were assigned to lower quantiles of the model (red and dark orange). Later on, when the demand returned to values closer to the climatology, the weights were re-adjusted. It can also be seen that in early 2010 the weights suffered a similar shock. Conversely, this corresponded to demand higher than the climatology (anomalously cold fall and winter), and therefore, the weights increased mainly for the upper quantiles of ECMWF (light oranges).

Rather than focus on any particular quantile of the distribution, one might want to evaluate the overall skill of the resulting mixtures. To address that, the average of the losses along Qgrid can be computed. It can be shown that provided that Qgrid is fine enough, this quantile mean pinball loss is a good approximation for the continuous ranked probability score (CRPS, Taieb et al., 2016).

Figure 23 presents the quantile-mean losses for all the experts, combinations and references; and the relative improvements of the combinations and models with respect to the uniform combination. It is seen that once again MLpol and BOA have the higher skill, closely followed by the oracle references and the MLpol_NWP and BOA_NWP combinations. All show improvements with respect to UNIF_NWP for lead weeks 1-4. These relative improvements (panel b) are on average around 5% for MLpol and BOA.

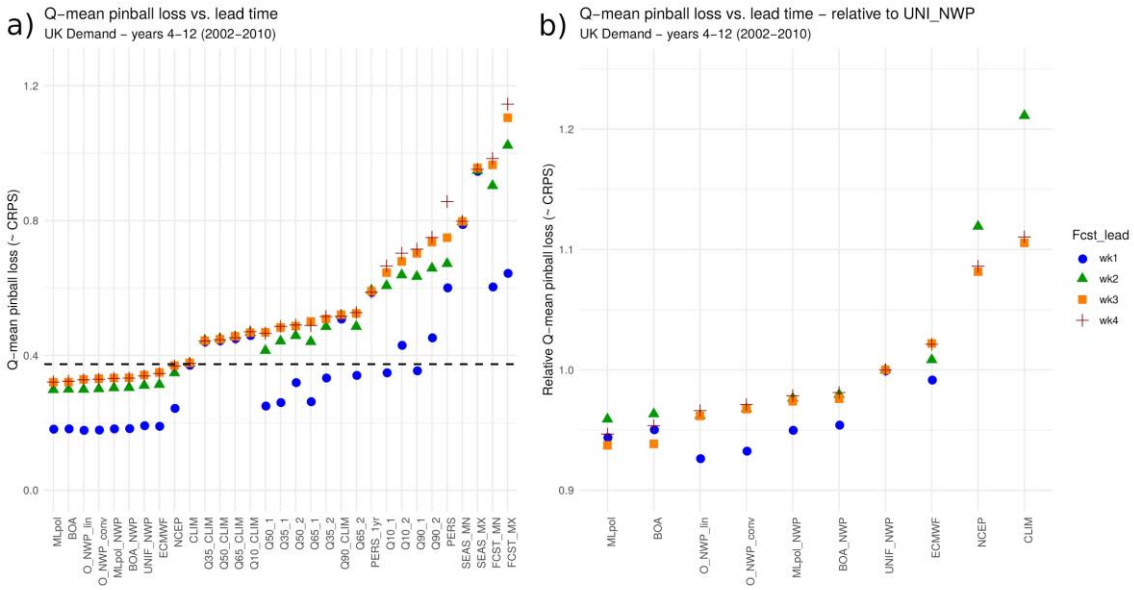


Figure 23. a) UK demand quantile-mean average pinball loss associated to the aggregation rules, the individual experts and the reference forecasts. The average losses for weeks 1 to 4 are presented as different symbols. The items on the x-axis are sorted from smaller to bigger loss based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for climatology, as a reference. b) Pinball losses for a subset of the models and references, but expressed as relative losses with respect to the UNIF_NWP combination, for each corresponding week.

These increases in skill, though moderate, seem quite robust. The statistical significance of the enhancements was assessed by using a Diebold-Mariano test to compare pairs of time series of quantile-mean losses. Results from those tests (not shown) revealed that MLpol and BOA resulted in forecasts that are significantly more skillful than UNIF_NWP for weeks 1-4. For lead weeks 2-4, the mixtures that included reanalysis-based experts (BOA, MLpol) were significantly more skillful than those that didn't (BOA_NWP, MLpol_NWP). MLpol gave predictions that were in general more skillful than those obtained from BOA, but not robustly across lead times.

6.4.4. Discussion and Conclusions

The innovative set of multi-model aggregation techniques presented here have shown very promising results for the enhancement of forecasting skill of country-level weekly demand beyond week 2, a horizon rarely beaten by the ECMWF system in the baseline skill assessment (S2S4E Deliverable 4.1, section 5.3.1).

MLpol and BOA showed significant skill improvements with respect to the climatology, standard multi-model methods and the individual best NWP system (ECMWF) for weeks 1-4. The increases in forecasting skill were of the order of 5% for lead weeks 3 and 4 with respect to a uniform combination, and between 7-12% with respect to the individual NWP systems. Part of this enhancement seems due to them including reanalysis-based experts. A case study illustrated how the algorithms are able to 'learn' from the performance of the experts through a season and adjust the weights accordingly to minimize losses. When reanalysis-based experts are included, this adjustment resulted in better predictions. Though it hasn't been included in

this summary, these methods were tested on other countries with large energy systems such as Germany, France and Spain, and the results presented here remained true.

7. Improving skill – Seamless S2S forecasting

7.1. Introduction

As sub-seasonal products became increasingly available, with the promise of improved skill for the month ahead, the need for efficient and optimal combination techniques between sub-seasonal and seasonal products appeared. Because S2S products are issued through different models with different initializations and boundary conditions, their added value is likely to vary in time and space (cf. Deliverable 4.1).

Previous work focusing on combining S2S products for hydrology in Europe have made the choice of systematically exploiting sub-seasonal forecasts up to 6 weeks ahead (Wetterhall and Di Giuseppe, 2018). Within the S2S4E Decision Support Tool, forecasts obtained from the S2S ECMWF products are used for different time horizons. When forecasting hydro-climate variables up to 4 weeks ahead, the forecast information from the sub-seasonal Extended Range product is displayed. When looking at horizons beyond the 4 weeks ahead, the information from the seasonal ECMWF SEAS5 product is displayed.

Here, we propose to have a closer look at the complementarity of the hydro-meteorological forecasts generated from the two ECMWF products. This investigation allows identifying optimal combination horizons, spatial variations in this optimal combination horizon, as well as sources for these differences in performance.

7.2. Methodology

7.2.1. Forcing post-processing

The ECMWF ER and SEAS5 precipitation and temperature reforecasts are first bias-adjusted against the HydroGFD corrected reanalysis (Berg et al., 2018). The resulting bias-adjusted reforecasts are produced at the 0.5 degree resolution grid of HydroGFD. These bias-adjusted precipitation and temperature reforecasts are then used to force the E-HYPE hydrological model (Donnelly et al., 2016; Hundecha et al., 2016; Lindström et al., 2010). In order to associate grid cells with each E-HYPE delineated catchment, the precipitation and temperature from the HydroGFD grid cell nearest to each catchment centroid are considered as representative of the entire upstream catchment area.

7.2.2. Hydrological model runs

The hydrological states of the E-HYPE model (e.g. lake levels, soil water content, snowpack) are initialized prior to each forecast run by forcing the model with the HydroGFD reanalysis up to the forecast issue date. Hydrological reforecasts of streamflow and snow are then produced for each forecast date based on the corresponding initial states and by forcing E-HYPE with the ECMWF ER and SEAS5 bias-adjusted precipitation and temperature. Following the forecast specifications, the hydrological runs based on the ECMWF SEAS5 forecasts are initialized on the 1st of each month and cover the following 7 months. The hydrological runs based on ECMWF ER forecasts are run only once a week (the product itself is initialized twice a week) and cover the following 6 weeks. All hydrological forecasts are produced at a daily time step.

7.2.3. Forecast evaluation

Hydrological forecasts based on the S2S meteorological forecasts are evaluated at the weekly time step based on the Continuous Rank Probability Score (CRPS; Hersbach, 2000). The CRPS is a probabilistic evaluation score that compares the forecast distribution with the step function corresponding to the observation (i.e. 0 for values smaller than the observation, and 1 elsewhere). This score penalizes biases, over-confidence and under-confidence, and therefore is often used to assess overall probabilistic performances. The reference used in the computation of the score is the perfect hydrological run (i.e. E-HYPE forced with HydroGFD precipitation and temperature). This type of reference was chosen because it highlights performances linked to the hydrological model forcing rather than performances of the hydrological model itself.

The performance of the systems based on ECMWF ER and SEAS5 in terms of CRPS is compared to a benchmark. The chosen benchmark is an ensemble based on E-HYPE historical runs corresponding to the forecast period but selected from years different from the one being forecast. The Continuous Rank Probability Skill Score (CRPSS hereafter) is computed by dividing the CRPS of the forecast systems by the CRPS of this benchmark. The values of the skill score range between minus infinity and 1, where 1 corresponds to a perfect forecast, 0 indicates that the benchmark and the system perform equally well, and negative values indicate that the forecast system has no skill with regard to the chosen benchmark.

7.2.4. Optimal combination horizon

We use the concept of optimal combination horizon, which we define here as the maximum horizon for which sub-seasonal forecasts are more skillful than the seasonal forecasts. Operationally, it corresponds to the horizon up to which there is added value from the sub-seasonal forecasts. The optimal combination horizon corresponds to the horizon week when a switch from sub-seasonal forecasts to seasonal forecasts yields the highest skill throughout the upcoming season.

7.3. A hydrological investigation – identification of critical lead times for S2S over Europe

7.3.1. When is the optimal combination horizon?

Despite no significant annual variations in terms of skill magnitude, the forecast dataset yielding the more skillful streamflow forecasts slightly varies with the time of year (Figure 24). In months such as April, July or October, streamflow forecasts based on the sub-seasonal product outperform the ones based on the seasonal product and will subsequently have, on average, skill further ahead. In other months, the sub-seasonal product performs very similarly to the seasonal forecasts. Overall, the skill magnitude is fairly close between both forcing products, and the main differences originate from the product specifications, i.e. their different horizons and their initialization frequencies.

Differences in skill appear when looking more closely at the skill pattern within the month (Figure 24). Firstly, the initialization frequency of the sub-seasonal product allows more skillful streamflow forecasts as early as in the second forecast date of the month. The later in the month the sub-seasonal forecast is issued, the greater the difference with the seasonal product

which is only initialized at the beginning of the month. Secondly, in some months such as January, February or September, the skill of the seasonal product decays more slowly than the skill of the sub-seasonal product. This is likely due to the differences in models producing the sub-seasonal and seasonal information. Indeed, the seasonal meteorological forecasts originate from a global climate model which accounts for slow climate phenomena driving predictability at longer time scales.

The analysis of the snow water equivalent forecasts yields very similar results in the months when snowpack exists in Europe. From November to April, the magnitude of the skill and its evolution with the lead time are similar to those obtained for streamflow forecasts. However, in October and November, during the snow accumulation season, sub-seasonal forecasts offer more frequently updated snow conditions, which seasonal forecasts do not provide. In the following, all snow forecast analyses are carried out only for issue month comprised between November and April.

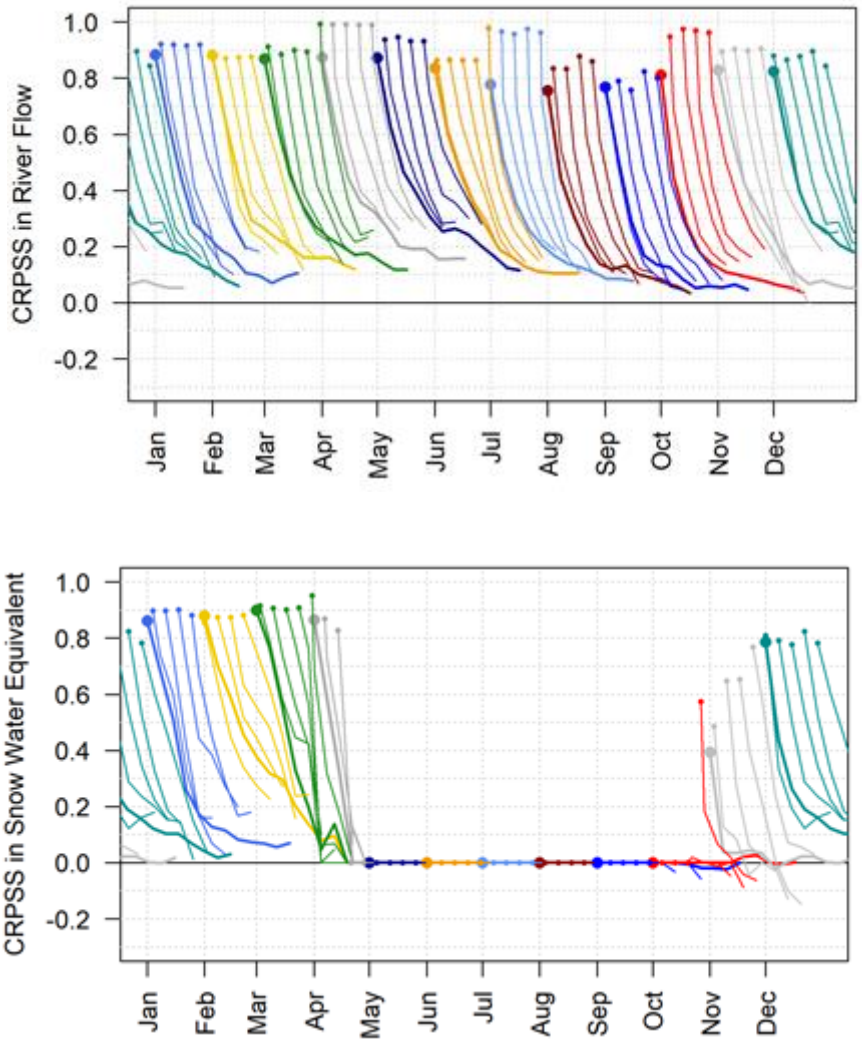


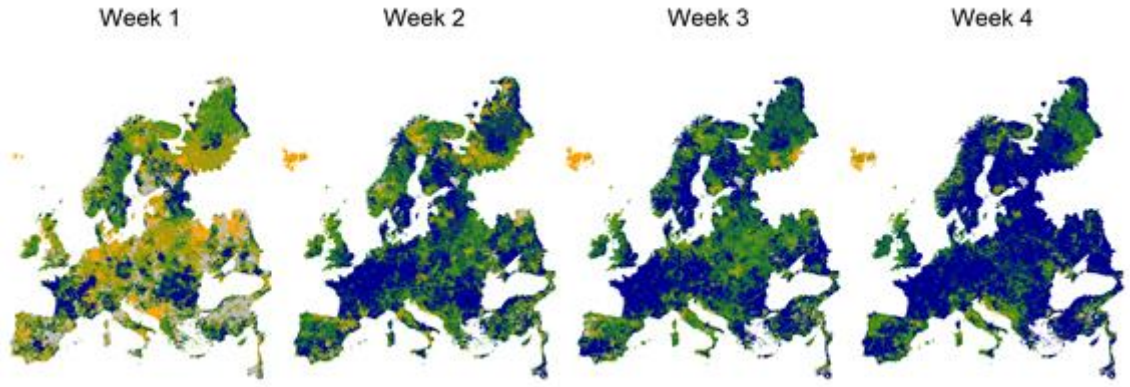
Figure 24. Median CRPSS in terms of streamflow (top) and snow water equivalent (bottom) over Europe per forecast initialisation date. Each colour corresponds to a calendar month. Thick lines correspond to the seasonal forecasts and thinner lines correspond to the sub-seasonal forecasts.

7.3.2. How does the optimal combination horizon vary spatially?

Results hint at the need for different combination horizons between sub-seasonal and seasonal forecasts depending on their issue dates and on the location. We observe in Figure 2 as well that the optimal combination horizon increases for later sub-seasonal forecast issue dates within the month (Figure 25). In the first week of the month, sub-seasonal forecasts already provide additional skill, with most regions attaining optimal combination horizons between one and 3 weeks ahead. Some well-defined regions, such as France, Romania and the Saint Petersburg area already benefit from sub-seasonal forecast up to 6 weeks ahead. The additional skill of sub-seasonal forecasts issued at the end of the month is more consistent with most values ranging between 4 and 6 weeks depending on the geographical region.

Spatial patterns in the optimal combination horizon can be identified (Figure 25). Regions with long combination horizons in terms of streamflow, indicating a greater input from sub-seasonal forecasts, are identified for example in Western Europe, along the western coast of the Black Sea and along the coasts of the Baltic Sea. In these catchments, the combination horizon for streamflow can extend up to 6 weeks as early as for the second issue week in the month.

The combination horizon for snow forecasts varies from 1 to 6 weeks depending on the region and the forecast issue week. Regions where sub-seasonal forecasts appear to have added value for snow, i.e. the optimal combination horizon is very short, include regions in Southern and Western Europe, possibly because they are less impacted by snow. In these regions, the skill of the sub-seasonal forecasts is limited to 1 to 2 weeks, even for forecasts issued later on in the month. Regions where the additional skill of sub-seasonal snow forecasts is very high, with optimal combination horizons up to 6 weeks as early as in the second issue week include Scandinavia, north-western Russia, Baltic countries and the Alps, which are regions known to be heavily impacted by snow.



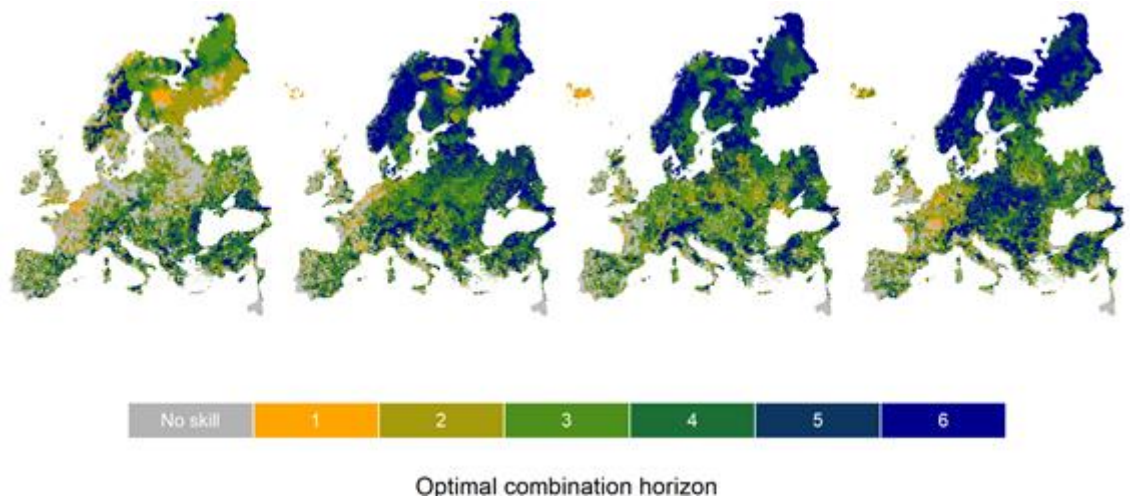


Figure 25. Maps of optimal combination horizons for streamflow (top) and snow water equivalent (bottom). Each column corresponds to a different sub-seasonal forecast issue week within the month.

7.3.3. Where does the additional skill of sub-seasonal forecasts come from?

We now analyze the cause for the additional skill found in streamflow and snow sub-seasonal forecasts, by analyzing the additional skill in precipitation and temperature forecasts. On the one hand, ECMWF ER precipitation forecasts generally have additional skill in comparison to seasonal forecasts up to 2 weeks ahead (Figure 26), and generally have significantly higher skill than ECMWF SEAS5 precipitation forecasts. On the other hand, ECMWF ER temperature forecasts exhibit longer optimal combination horizons up to 3 to 4 weeks (Figure 26). The magnitude of the skill in temperature however, is very close to that of seasonal forecasts. The difference in optimal combination time between precipitation and temperature forecasts and streamflow and snow forecasts highlights the non-linearity of the natural processes linking meteorological and hydrological variables. Gains up to one to 4 weeks in terms of meteorological forecasts can lead to gains up to 6 weeks in terms of hydrological forecasts, simply because streamflow and snow processes have a longer memory than the atmosphere does.

A comparison of the spatial patterns of optimal combination horizons in precipitation and temperature (Figure 26) and in streamflow and snow (Figure 25) does not show clear correspondences between meteorological and hydrological forecasts. This is likely due to the different spatial units in meteorology and hydrology, e.g. precipitation upstream the river will have an impact on all its downstream area.

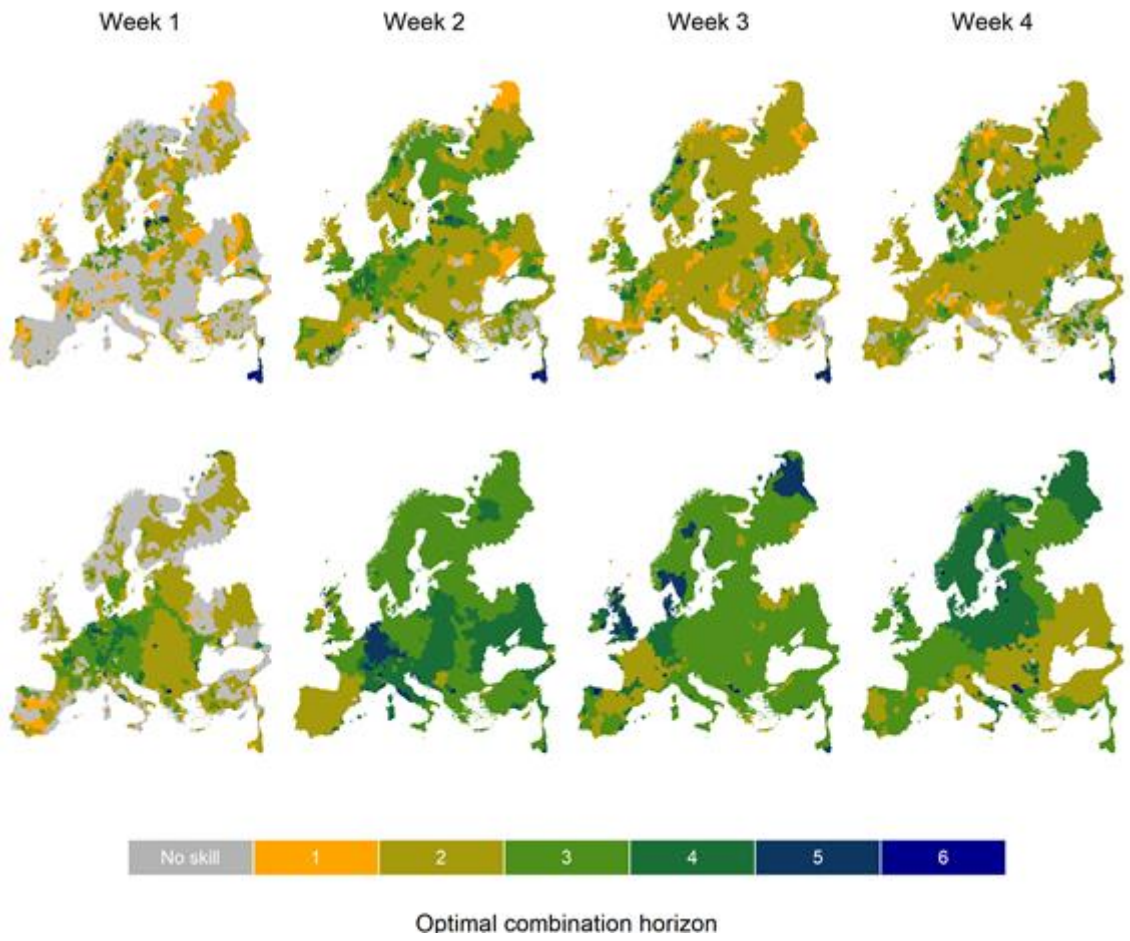


Figure 26. Maps of optimal combination horizons for precipitation (top) and temperature (bottom). Each column corresponds to a different sub-seasonal forecast issue week within the month.

7.4. Discussion and Conclusions

The results from the analysis can be summarized in general guidance on how to combine sub-seasonal and seasonal forecasts for streamflow and snow forecasting. For streamflow forecasts, two regions were identified:

- In mountainous and rain-driven regions (grey regions in Figure 27a), optimal combination horizons range between 2 weeks ahead (when the forecast is issued at the beginning of the month) and 4 weeks ahead (when the forecast is issued at the end of the month).
- In most European catchments (red regions in Figure 27a), sub-seasonal forecasts are skillful on average up to 3 weeks ahead when the forecast is issued in the first week of the month, and up to 6 weeks ahead for other issue weeks.

For snow forecasts, three regions were identified (Figure 27b).

- Most regions (lighter shade of red in Figure 27b) benefit from sub-seasonal forecasts up to 4 weeks ahead, as early as in the second issue week of the month.
- In regions heavily impacted by rainfall and snowfall (darker shade of red in Figure 27b), the optimal combination horizon is longer than in other regions. The forecasts issued in

the first week of the months have optimal combination horizons of about 2 weeks. Other forecasts have combination horizons of about 5 weeks.

- In some regions of southern Europe, however, the optimal combination horizon is limited to 2 to 3 weeks ahead only from the second forecast issue week (grey areas in Figure 27b), probably because the impact of snow in these regions is limited and therefore no clear differences can be identified between the seasonal and sub-seasonal forecasts.

The main conclusions from this investigation focusing on snow and streamflow are the following:

- The level of skill of sub-seasonal and seasonal forecasts is comparable in terms of magnitude for most months of the year, suggesting that the added value of sub-seasonal forecasts comes from frequent initializations rather than from improved precipitation and temperature forecasts.
- Results hint at the need for different combination horizons between sub-seasonal and seasonal forecasts depending on their issue dates and location:
  - The optimal combination horizon increases for sub-seasonal forecast issued from the second week of the month.
  - Regions with long combination horizons in terms of streamflow were identified in Western Europe, along the western coast of the Black Sea and along the coasts of the Baltic Sea.
  - Regions with long combination horizons in terms of snow were identified in Scandinavia, north-western Russia, Baltic countries and the Alps.
- Gains in meteorological skill from sub-seasonal forecasts (1 to 3 weeks) are amplified, both in terms of magnitude and maximum horizon with skill, when looking at downstream hydrological processes such as streamflow and snow (often up to 6 weeks).

(a) (b)

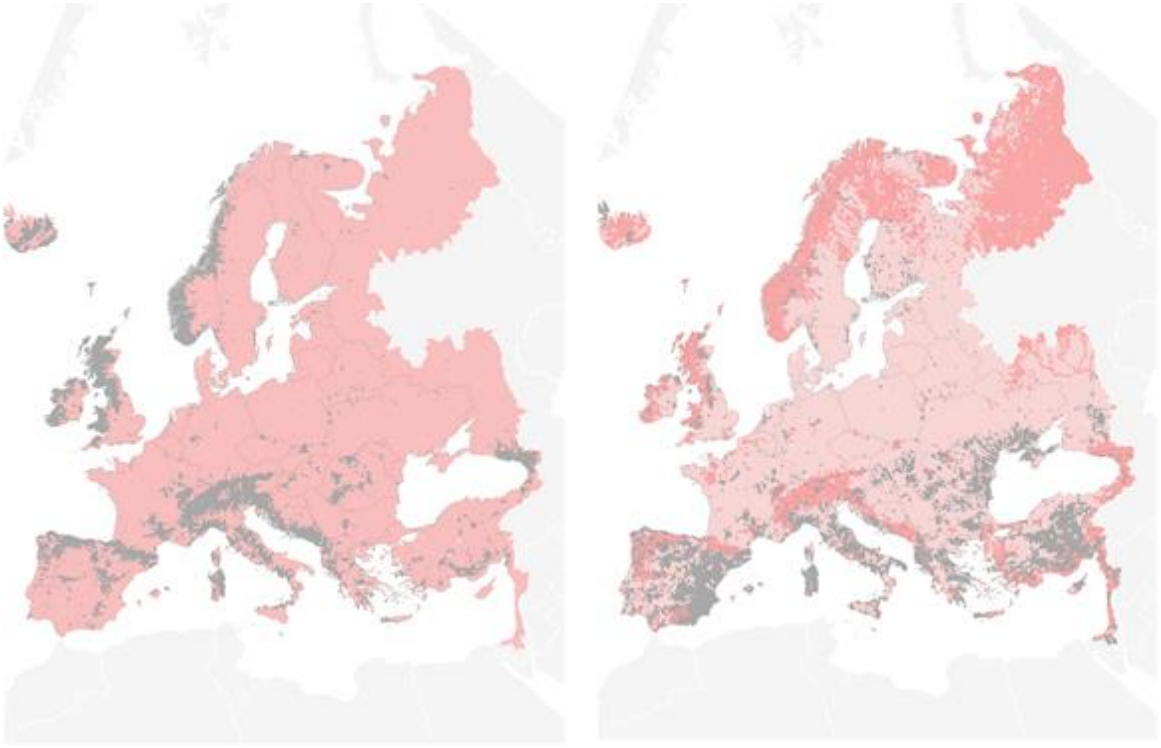


Figure 27. Skill regions identified over Europe for (a) streamflow and (b) snow water equivalent (right). Regions where sub-seasonal forecasts have a longer combination horizon appear in shades of red, others appear in grey.

8. Conclusions

This report describes a set of different state-of-the-art approaches and methodologies that could potentially improve the forecasting skill of various variables at time horizons from sub-seasonal to seasonal. In addition, this report builds on insights and recommendations provided in deliverable D4.1 and highlights results and conclusions from the investigations conducted in deliverables D4.2 and D4.3. Consequently it gives an overall view of the science conducted in work package 4 of the S2S4E project. The conclusions from this report are clustered depending on the specific investigation. Note that here only the key messages are provided, whilst a detailed presentation is given in the corresponding chapters of this report.

The investigation conducted using *pattern-based techniques* indicates that most of the scientific developments (those also referred in deliverables D4.2 and D4.3) are not straightforward to be implemented in existing climate services (i.e. the S2S4E DST) which are specifically designed to provide probabilistic forecast information from dynamic models. Although this is achievable, incorporation of the techniques in existing tools can be confusing to users, whilst training would be needed for better understanding of the science behind the techniques. As an example for service evolution, there could be showcases highlighting the added-value in the seasonal outlooks showing forecasts of NAO teleconnections.

The analysis using the *lagged ensembles and calibration* showed that the impact of the number of lagged ensemble members (i.e. 4, 8 and 12 members) on the forecasting skill is not significant; only the lagged ensemble of 12 members indicated a slight improvement in skill for the short lead weeks and in specific locations (equatorial region, Atlantic and Indian Ocean). In addition, the investigation showed that the calibration of the temperature forecasts can reduce the negative skill in some areas, whilst the sample of calibration data can affect the skill; with a large sample of 20 years achieving a higher skill than a sample of 12 years. This specific conclusion was highlighted in the mid latitudes in both hemispheres, where areas that had negative forecasting skill with the 12-year calibration sample, achieved positive skill with the 20-year calibration sample. Finally, the investigation showed that the calibration window can affect the calibration results. Here, the analysis concluded that a 3-week window is optimal for the calibration of the sub-seasonal forecasts leading to robust climatology estimation and an acceptable skill.

A set of three different *bias-adjustment methods* were tested on how they can affect the forecasting skill. The added-value from the bias-adjustment method depends on the variable of interest with temperature generally showing higher skill than precipitation. In addition, the temporal resolution of the adjustment is another factor controlling the skill, with a monthly adjustment achieving higher skill than a daily adjustment. The analysis also concluded that when skill is already present in the raw forecasts, all three bias-adjustment methods managed to further improve the skill (i.e. monthly adjustment of temperature). However when there is no (or low) skill in the raw forecasts, the bias-adjustment methods can improve the skill but still without reaching positive values (i.e. daily adjustment of precipitation). Finally, the investigation did not identify a bias-adjustment method that was superior to another.

The investigations focusing on *multi-modeling approaches* highlighted the high potential of this field. The solar radiation and temperature analysis using three different climate models (i.e. ECMWF, MF, DWD) available in the Climate Data Store of the C3S highlighted the importance of snow-albedo processes for prediction in winter and the effect of the atmospheric dynamics in

summer. The analysis focusing on the seasonal predictability of streamflow highlighted the dependency of the skill both geographically and seasonally with higher skill being observed over Europe in summer and autumn in comparison to winter and spring. A multi-model (ECMWF, MF and UKMO) averaged product based on simple equal model weighting indicated improvements in the forecasting skill even when the individual climate models did not show skill. This was observed in high lead times. Here the multi-model averaged product showed improvement in the spatial variability of the skill over the entire domain with more regions achieving positive skill than under the individual models. The multi-model approach finally showed to be promising for the country-level weekly demand, particularly beyond week 2. The two methodologies (MLpol and BOA) showed significant improvements with respect to climatology, standard multi-model methods and the individual best NWP system (ECMWF) for weeks 1-4.

Last but not least, the investigation on *seamless sub-seasonal to seasonal forecasting* identified the critical lead times for combining the S2S forecasts and also the sources of skill in the sub-seasonal forecasts. The added value of sub-seasonal streamflow forecasts comes from the frequent initializations of the hydrological model rather than from improved precipitation and temperature forecasts. In addition, the gains in the meteorological skill from sub-seasonal forecasts are amplified both in terms of magnitude and maximum horizon with skill, when looking at streamflow and snow. This investigation set a guide to users on how to combine S2S forecasts for streamflow and snow. For streamflow and snow, two and three regions were identified respectively where the optimal combination horizons of S2S forecasts varies. This is a very interesting conclusion since the investigation can guide the users on selecting either the sub-seasonal or seasonal product depending on their region and initialization month.

Bibliography

Alessandri A., De Felice M., Catalano F., Lee J.-Y., Wang B., Lee D. Y., Yoo J.-H., Weisheimer A., 2018: Grand European and Asian-Pacific multi-model seasonal forecasts: maximization of skill and of potential economical value to end-users. Clim. Dyn. 50, 2719-2738, doi:10.1007/s00382-017-3766-y

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

Arnal, L., A. W. Wood, E. Stephens, H. L. Cloke, and F. Pappenberger (2017), An Efficient Approach for Estimating Streamflow Forecast Skill Elasticity, J. Hydrometeorol., 18, 1715–1729, doi:10.1175/JHM-D-16-0259.1.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., … Pappenberger, F. (2018). Skilful seasonal forecasts of streamflow over Europe ? Hydrology and Earth System Sciences, 22, 2057–2072. https://doi.org/doi.org/10.5194/hess-22-2057-2018

Bartók B., I. Tobin I., R. Vautard, M. Vrac, X. Jin, G. Levavasseur, S. Denvil, L. Dubus, S. Parey, P-A. Michelangeli, A. Troccoli, Y-M. Saint-Drenan, 2019, A climate projection dataset tailored for the European energy sector, Climate Services, Volume 16, 100138, ISSN 2405-8807,https://doi.org/10.1016/j.cliser.2019.100138.

Bennett, J. C., Wang, Q. J., Robertson, D. E., Schepen, A., Li, M., & Michael, K. (2017). Assessment of an ensemble seasonal streamflow forecasting system for Australia. Hydrology and Earth System Sciences, 21(12), 6007–6030. https://doi.org/10.5194/hess-21-6007-2017

Berg, P., Donnelly, C., Gustafsson, D., 2018. Near-real-time adjusted reanalysis forcing data for hydrology. Hydrology and Earth System Sciences 22, 989–1000. https://doi.org/10.5194/hess-22-989-2018

Bierkens, M. F. P., & Van Beek, L. P. H. (2009). Seasonal predictability of european discharge: NAO and hydrological response time. Journal of Hydrometeorology, 10(4), 953–968. https://doi.org/10.1175/2009JHM1034.1

Bloomfield, H. C., Brayshaw, D. J., & Charlton-Perez, A. J. (2019). Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types. Meteorological Applications.

Bruno Soares, M., Alexander, M., & Dessai, S. (2017). Sectoral use of climate information in Europe: A synoptic overview. Climate Services, 1–16. https://doi.org/10.1016/j.cliser.2017.06.001

Bruno Soares, M., & Dessai, S. (2016). Barriers and enablers to the use of seasonal climate forecasts amongst organisations in Europe. Climatic Change, 137(1–2), 89–103. https://doi.org/10.1007/s10584-016-1671-8

Cesa-Bianchi, N., & Lugosi, G. (2006). Prediction with Expert Advice. In Prediction, Learning, and Games (pp. 7-39). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511546921.003

Crochemore, L., Ramos, M.-H., & Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. Hydrology and Earth System Sciences, 20, 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

DelSole, T., Trenary, L., Tippett, M. K., & Pegion, K. (2017). Predictability of week-3-4 average temperature and precipitation over the contiguous United States. *Journal of Climate*, *30*(10), 3499–3512. https://doi.org/10.1175/JCLI-D-16-0567.1

DelSole, T., 2007: A Bayesian Framework for Multimodel Regression. J. Climate, 20, 2810–2826, https://doi.org/10.1175/JCLI4179.1

Demirel, M. C., Booij, M. J., & Hoekstra, A. Y. (2013). Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models. Water Resources Research, 49(7), 4035–4053. https://doi.org/10.1002/wrcr.20294

Doblas-Reyes, F. J., Hagedorn, R., Palmer, T. N. N., Doblas-Reyes, F. J., Palmer, T. N. N., HAGEDORN, R., & Palmer, T. N. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination By. Tellus A: Dynamic Meteorology and Oceanography, 57(3).

Donnelly, C., Andersson, J.C.M., Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. Hydrological Sciences Journal 61, 255–273. https://doi.org/10.1080/02626667.2015.1027710

ENTSO (2019). European network of transmission system operators for electricity: data platform [cited 17th Jan 2017]. Available from http://entsoe.eu/data/Pages/default.aspx.

Gaillard, P., Stoltz, G., & Van Erven, T. (2014). A second-order bound with excess losses. In COLT. arXiv:1402.2044.

Global Wind Atlas (2019). The Global Wind Atlas [cited 1st March 2019]. Available from https://globalwindatlas.info.

Green, R. (2005). Electricity and Markets. Oxford Review of Economic Policy, 21, 67–87.

Greuell, W., Franssen, W. H. P., Biemans, H., & Hutjes, R. W. A. (2018). Seasonal streamflow forecasts for Europe – Part I : Hindcast verification with pseudo- and real observations. Hydrology and Earth System Sciences, 22, 3453–3472. https://doi.org/10.5194/hess-22-3453-2018

Greuell, W., Franssen, W. H. P., & Hutjes, R. W. A. (2019). Seasonal streamflow forecasts for Europe - Part 2: Sources of skill. Hydrology and Earth System Sciences, 23(1), 371–391. https://doi.org/10.5194/hess-23-371-2019

Hagedorn R., Doblas-Reyes F.J., Palmer, T.N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus Ser A Dyn Meteorol Oceanogr 57(3):219–233. doi:10.1111/j.1600-0870.2005.00103.x

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather and Forecasting 15, 559–570.

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

Hersbach, H, Bell, W, Berrisford, P, Horányi, A, J., M-S, Nicolas, J, Radu, R, Schepers, D, Simmons, A, Soci, C, Dee, D (2019) Global Reanalysis: goodbye ERA-Interim, hello ERA5. ECMWF Newsletter, p 17-24. http://dx.doi.org/10.21957/vf291hehd7

Hundecha, Y., Arheimer, B., Donnelly, C., Pechlivanidis, I., 2016. A regional parameter estimation scheme for a pan-European multi-basin model. Journal of Hydrology: Regional Studies 6, 90–111. https://doi.org/10.1016/j.ejrh.2016.04.002

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

Kim, H. M., Webster, P. J., & Curry, J. A. (2012). Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere Winter. Climate Dynamics, 39(12), 2957–2973. https://doi.org/10.1007/s00382-012-1364-6

Klein, B., D. Meissner, H. U. Kobialka, and P. Reggiani (2016), Predictive uncertainty estimation of hydrological multi-model ensembles using pair-copula construction, Water, 8(4), 1–22, doi:10.3390/w8040125.

Li, H., Luo, L., Wood, E. F., & Schaake, J. (2009). The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. Journal of Geophysical Research, 114(D4), D04114. https://doi.org/10.1029/2008JD010969

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., Arheimer, B., 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrol Res 41, 295. https://doi.org/10.2166/nh.2010.007

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

Lledó, Ll. (2017). CLIM4ENERGY technical note no.1: computing capacity factor. BSC-ESS Technical Memorandum 2017-001, 9 pp.

Ll. Lledó, V. Torralba, A. Soret, J. Ramon, F.J. Doblas-Reyes, 2019, Seasonal forecasts of wind power generation, Renewable Energy, Volume 143, Pages 91-100, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2019.04.135.

Manzanas, R., Gutiérrez, J.M., Bhend, J., Hemri, S., Doblas-Reyes, F.J., Torralba, V., Penabad, E., Brookshaw, A. (2019). Bias adjustment and ensemble recalibration methods for seasonal forecasting: a comprehensive intercomparison using the C3S dataset. Climate Dynamics: 1-19, doi:10.1007/s00382-019-04640-4.

Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., … Arnold, J. R. (2017). An intercomparison of approaches for improving operational seasonal streamflow forecasts. Hydrology and Earth System Sciences, 21(7), 3915–3935. https://doi.org/10.5194/hess-21-3915-2017

Michelangeli, P-A., M. Vrac, and H. Loukos, 2009, Probabilistic downscaling approaches: Application to wind cumulative distribution functions, Geophys. Res. Lett., 36, L11708, doi:10.1029/2009GL038401.

Muhammad, A., T. A. Stadnyk, F. Unduche & P. Coulibaly (2018): Multi-Model Approaches for Improving Seasonal Ensemble Streamflow Prediction Scheme with Various Statistical Post-Processing Techniques in the Canadian Prairie Region. 10(11), 1604

Raftery, A. E., T. Gneiting, F. Balabdaoui & M. Polakowski (2005): Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 133(5), 1155-1174

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., … Becker, E. (2014). The NCEP Climate Forecast System Version 2. Journal of Climate, 27(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Sansom, P. G., D. Stephenson, C. Ferro, G. Zappa, and L. Shaffrey, 2013: Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. J. Climate, 26, 4017–4037.

Shukla, S., & Lettenmaier, D. P. (2011). Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. Hydrology and Earth System Sciences, 15(11), 3529–3538.

Siebert, S., DB Stephenson (2019). 'Forecast Recalibration and Multimodel Combination', in Robertson, A., F. Vitart (ed.) S2S Prediction: The Gap Between Weather and Climate Forecasting. Elsevier, pp. 321-336. https://doi.org/10.1016/B978-0-12-811714-9.00015-2

Stoft, S. (2002). Power System Economics. IEEE Press Wiley, Piscataway, NJ.

Taieb S.B., R. Huser, R. J. Hyndman and M. G. Genton (2016) Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression, in IEEE Transactions on Smart Grid, vol. 7, no. 5, pp. 2448-2455. doi: 10.1109/TSG.2016.2527820

Thiboult, A., Anctil, F., & Boucher, M. A. (2016). Accounting for three sources of uncertainty in ensemble hydrological forecasting. Hydrology and Earth System Sciences, 20(5), 1809–1825. https://doi.org/10.5194/hess-20-1809-2016

Torralba, V., Doblas-Reyes, F. J., MacLeod, D., Christel, I., & Davis, M. (2017). Seasonal Climate Prediction: A New Source of Information for the Management of Wind Energy Resources. Journal of Applied Meteorology and Climatology, 56(5), 1231–1247. https://doi.org/10.1175/JAMC-D-16-0204.1

Trenary, L., Delsole, T., Tippett, M. K., & Pegion, K. (2017). A new method for determining the optimal lagged ensemble. Journal of Advances in Modeling Earth Systems, 9, 291–306. https://doi.org/10.1002/2016MS000838

Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., & Beck, H. E. (2013). Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide. Water Resources Research, 49(5), 2729–2746. https://doi.org/10.1002/wrcr.20251

Vrac, M., Noël, T., Vautard, R., 2016, Bias correction of precipitation through Singularity Stochastic Removal: because occurrences matter. J. Geophys. Res. 121, 5237–5258.

Vrac, M., Drobinski, P., Merlo, A., Herrmann, M., Lavaysse, C., Li, L., Somot, S., 2012. Dynamical and statistical downscaling of the French Mediterranean climate: uncertainty assessment. Nat. Hazards Earth Syst. Sci. 12, 2769–2784.

Wanders, N., and E. F. Wood (2016), Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations, Environ. Res. Lett., 11(9), 094007, doi:10.1088/1748-9326/11/9/094007.

Wanders, N., S. Thober, R. Kumar, M. Pan, J. Sheffield, L. Samaniego, and E. F. Wood (2019), Development and evaluation of a Pan-European multi-model seasonal hydrological forecasting system, J. Hydrometeorol., 20, 99–115, doi:10.1175/JHM-D-18-0040.1.

Wang, L., Robertson, A. W. (2019). Week 3–4 predictability over the United States assessed from two operational ensemble prediction systems. *Climate Dynamics*, *52*(9–10), 5861–5875. https://doi.org/10.1007/s00382-018-4484-9

Wetterhall, F., Di Giuseppe, F., 2018. The benefit of seamless forecasts for hydrological predictions over Europe. Hydrology and Earth System Sciences 22, 3409–3420. https://doi.org/10.5194/hess-22-3409-2018

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

Wintenberger, O. (2017) Optimal learning with Bernstein online aggregation. Mach Learn, 106: 119. https://doi.org/10.1007/s10994-016-5592-6

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & Clark, M. (2016). Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. Journal of Hydrometeorology, 17(2), 651–668. https://doi.org/10.1175/JHM-D-14-0213.1

Wood, A. W., & Lettenmaier, D. P. (2008). An ensemble approach for attribution of hydrologic prediction uncertainty. Geophysical Research Letters, 35(14), 1–5. https://doi.org/10.1029/2008GL034648

Yang, W., Andréasson, J., Graham, P. L., Olsson, J., Rosberg, J., & Wetterhall, F. (2010). Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. Hydrology Research, 41(3–4), 211–229. https://doi.org/10.2166/nh.2010.004

Yossef, N. C., Winsemius, H., Weerts, A., Van Beek, R., & Bierkens, M. F. P. (2013). Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. Water Resources Research, 49(8), 4687–4699. https://doi.org/10.1002/wrcr.20350

Yossef, N., Van Beek, R., Weerts, A., Winsemius, H., & Bierkens, M. F. P. (2017). Skill of a global forecasting system in seasonal ensemble streamflow prediction. Hydrology and Earth System Sciences, 21(8), 4103–4114. https://doi.org/10.5194/hess-21-4103-2017

Zhang, X., Tang, Q., Leng, G., Liu, X., Li, Z., & Huang, Z. (2017). On the dominant factor controlling seasonal hydrological forecast skill in China. Water, 9(902), 1–19. https://doi.org/10.3390/w9110902

Zhao, T., J.C. Bennett, Q.J. Wang, A. Schepen, A.W. Wood, D.E. Robertson, and M. Ramos, 2017, How Suitable is Quantile Mapping For Post-processing GCM Precipitation Forecasts? J. Climate, 30, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1

Annex 1 - The impacts of using midnight vs. six hourly wind speeds to create wind power capacity factors for the S2S4E Decision Support Tool

Context and summary

In order to calculate weekly-mean wind speeds and wind power generation for the S2S4E decision support tool (DST) the extended range forecast 10m wind speed data is used (currently NCEP CFS, proposed to update to ECMWF via S2S database). A potential problem with this is that the 10m wind speeds are only available as midnight values. The ERA5 based wind power model used in S2S4E has been developed using six hourly wind speed data, and the ECMWF seasonal forecasting model outputs used in the DST are six-hourly data, both of which are high enough resolution to represent the diurnal cycle of wind speed. The sub-seasonal forecast to be used operationally in the DST, however, cannot represent diurnal cycles due to the coarse temporal sampling.

The outputs shown on the DST are weekly-averages, so it is important to understand what impact this limited temporal resolution has on the DST weekly-mean values for wind, wind-power and country-aggregates such as wind power and demand-net-wind. This document investigates this by artificially downgrading the ERA5 to daily values (rather than 6-hourly) and comparing the resulting time-series to those calculated from the same dataset but using 6-hourly data. To set this difference in performance in context, the country-aggregate values produced are also compared to the difference between the "best" ERA5 method (6-hourly) and actual observed country-level power system data.

The analysis in this report suggests that the use of daily (midnight) wind speeds has a quantitative impact compared to using higher-frequency (6hourly) data, particularly in the summer season (local differences of up to ~1m/s or 5% CF depending on turbine type, equivalent to up to ~few %CF at country level). The variability in the resulting weekly time-series remains, however, very strongly correlated (i.e. if the wind power output above or below average is well-represented). Moreover, at country-aggregate level, the CF differences between the two different data sources (daily vs 6hourly) are, at worst, no greater than those between the "best" method (6hourly) and a corresponding set of direct power system observations. This therefore suggests that while the net effect of using daily (rather than 6-hourly) data is detrimental to performance, this impact is modest when the existing performance limitations of the DST are considered.

It is therefore recommended that it be clearly explained to potential users of the DST that the absolute values (rather than anomalies) of variables taken from the DST be used with some caution (particularly in summer, when the impact of the lower temporal resolution is greatest and forecast skill is lower than in winter). This is, however, a general limitation that should already be being communicated to DST users irrespective of the time-frequency of the sub-seasonal forecast data being used.

Background: The wind power model

The S2S4E grid point wind power model works by taking bias corrected 100m wind speeds and passing them through a set of three different classes of wind turbines shown in Figure 1. Each turbine has a preferable set of wind speeds to operate within. The country aggregate wind power model takes this process one step further, by calculating the long-term average wind speed in each grid box and then assigning the most appropriate turbine for these long-term wind conditions. The grid point wind power generation is then weighted by where known turbines are installed and aggregated to country level. For more information on the modelling framework see Deliverables 3.2, 4.1, 4.2, 4.3. This information is important to bear in mind because differences between weekly- mean wind power generation using midnight or six-hourly data will be most important in the regions where we know large amounts of turbines are installed.

When working with the forecasts datasets the same procedure is applied. However, the 100m wind speed data is first calibrated to the bias corrected ERA5 reanalysis using a leave-one-out lead time dependent bias correction on the mean and variance of the data. As only the midnight winds are available from the forecasts only the midnight winds are used in the forecast calibration.

Bias corrected 100m wind speeds

This sections analyses the differences between weekly-mean bias corrected 100m wind speeds calculated using midnight wind speeds vs. six-hourly wind speeds from ERA5. This data is used as it is the data which the forecasts are calibrated to. ERA5 has also been degraded to a 1.5 degree grid in order to compare most favourably with the forecasts used in the DST.

Figure 2 shows composites of weekly-mean wind speed from 1980-2018 for March, June, September and December (the first month of each meteorological season) created from the midnight 100m wind speeds and the six-hourly 100m wind speeds. A seasonal cycle is present with highest winds seen in winter and over the North Atlantic storm track region. The difference plots between the two sets of weekly-means show that the largest differences are seen in summer (-0.5ms^{-1} over most of Europe, except for over Norway and the Mediterranean where +0.5ms^{-1} differences are seen.) The negative differences over most of Europe in summer show that the midnight wind speeds are higher than those found in day-time hours, which could lead to errors in the amount of wind power generation.

The correlation between the weekly-mean 100m wind speeds generated from the midnight and six- hourly data are also shown in Figure 2. There are >0.9 for most of the year, except for some values from 0.7-0.9 over Southern Europe in March and September. In summer the areas of largest wind speed differences are associated with reduced correlation. However these values do not drop below 0.6 and are generally high in regions where wind farms are installed (see Figure 1) except for in Spain.

Wind power capacity factor: the grid point model

Figure 3 shows the differences between average weekly-mean capacity factor for March, June, September and December for the three turbines shown in Figure 1. A similar pattern is seen as in Figure 2 for the 100m wind speed data, the largest differences are seen in over Europe in summer, and they are most pronounced for the smallest turbine (the Vestas) which is most sensitive to changes at lower wind speeds often present over land (see the power curves in Figure 1).

Differences are up to ~5% in summer with some of the largest differences seen in France and Spain. In the North Sea region where large amounts of European offshore wind is installed there are only very small differences for all of the turbines.

Figure 4 shows the correlation between the weekly-mean capacity factors created from the midnight wind speeds and six-hourly wind speeds. As seen for the 100m wind speeds, in all seasons correlation greater than 0.6 is seen over Europe with the highest correlation seen in winter (0.9) and lowest in summer (0.6). The lowest correlations are seen in the extended winter season over the North Atlantic. On further investigation this is thought to be due to the extremely variable six-hourly wind speeds over this region with it being common to exceed the wind turbine cut out speeds of $20ms^{-1}$ (for the Vestas) and $25ms^{-1}$ (for the Enercon and Gamesa) turbines which will lead to a drop to zero in capacity factor for that hour. It is important to note these are not regions where turbines are installed and the correlation between the two versions of the weekly-mean wind speeds is still at worst 0.6.

These results show that although there are differences between the weekly-mean wind speeds calculated from midnight vs. six-hourly data, the two versions of the weekly-means are very highly correlated. This suggests that the information provided from the forecasts could still be useful to the users. The time of year where the differences are smallest (extended winter) is also the time of year where forecast skill is highest, so the limitation of midnight wind speeds should not be having too much of a negative impact. Caution should be taken in summer if users wish to use exact values from the DST as the midnight winds produce an over-estimate.

Wind power capacity factor: the country-level model

The previous section highlighted that the areas of largest differences when comparing weekly-mean capacity factors generated using the midnight vs. six-hourly wind speeds are in summer, and located over Spain France and Germany. This season and these countries are therefore focused on in this section.

Figure 5 shows the histograms comparing the weekly-mean country-level capacity factor for summer (June, July and August) generated using the two temporal resolutions of data. The distributions are generally quite similar with a mean difference of ~2% in each country. Very high correlation is seen for the country-level model (>0.9) for the weekly-mean wind speeds created using midnight vs. six- hourly wind speeds.

An explanation as to why the largest differences may be present in summer is given in the right-most sub-panels of Figure 5. This shows the mean diurnal cycles of country-level capacity factor for the year, winter and summer. The largest diurnal cycles are seen in summer for France and Spain. The largest differences between the daily mean wind speed and the midnight wind speed are also seen in summer.

At a country-level the differences between weekly-mean capacity factors are relatively small. This suggests that the country level models could still provide useful information to energy users.

Weekly-mean differences in context

This document has shown that differences seen between the weekly-mean capacity factors derived using midnight vs. six hourly winds are relatively small for the majority of the year except for in summer. In this section we put these differences into the wider context of the modelling framework. Figure 6 shows comparisons of the ERA5 1.5 degree data used in the

previous analysis to the native- resolution ERA5 wind power model (used in work package 3 deliverables) and then compared to the 2018 ENTSOe data. France, Spain and Germany are kept as case study countries for continuity.

In France there are large differences between the ERA5 model at native resolution (~0.3 degrees) compared to the 1.5 degree model, which are much larger than the differences between the midnight and six-hourly wind speeds. In the case for France, the native resolution ERA5 model is much closer to the observed capacity factors. Similar behavior is seen for Spain, where the difference between the re-analysis spatial resolutions leads to greater differences than the change in model temporal resolution. In this case the 1.5 degree model looks slightly closer to the observations. In Germany the differences between all of the re-analysis derived models are relatively small compared to the difference between the re-analysis models and observations.

It can therefore be concluded that the differences caused by the different temporal resolution are comparable or less (depending on the chosen country) than the differences caused by the degradation of ERA5's spatial resolution in order to be compatible with the sub-seasonal forecasting gridded output.

Conclusion

The analysis in this report suggests that even though only the midnight wind speeds are available from the sub-seasonal forecasts output these can still provide useful information to users interested in the relative changes (i.e. is the wind power output above or below average for this region).

However, caution should be taken if using the absolute values from the DST, particularly in summer when the largest differences are seen and forecast skill is also at its lowest.

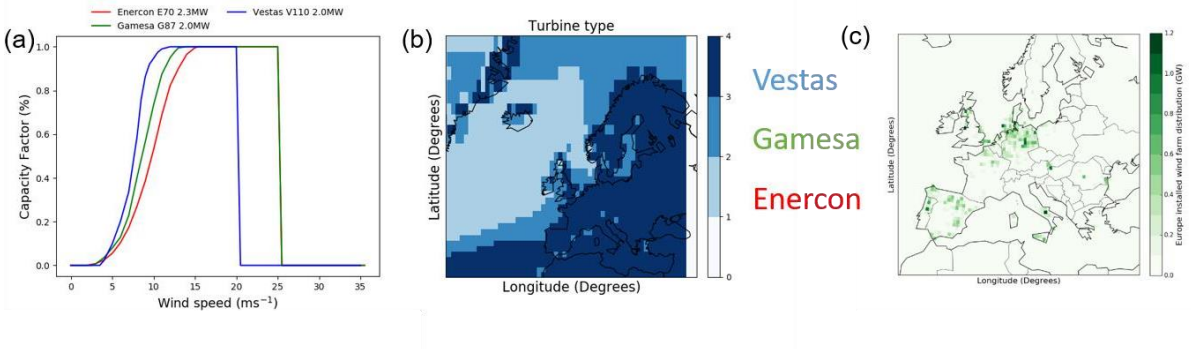


Figure A 1. (a) The three types of wind turbine used in this study (b) The most appropriate turbine installed in each grid box based on the ERA5 1980-2018 mean wind speed (c) The location of the 2017 installed wind power generation across Europe.

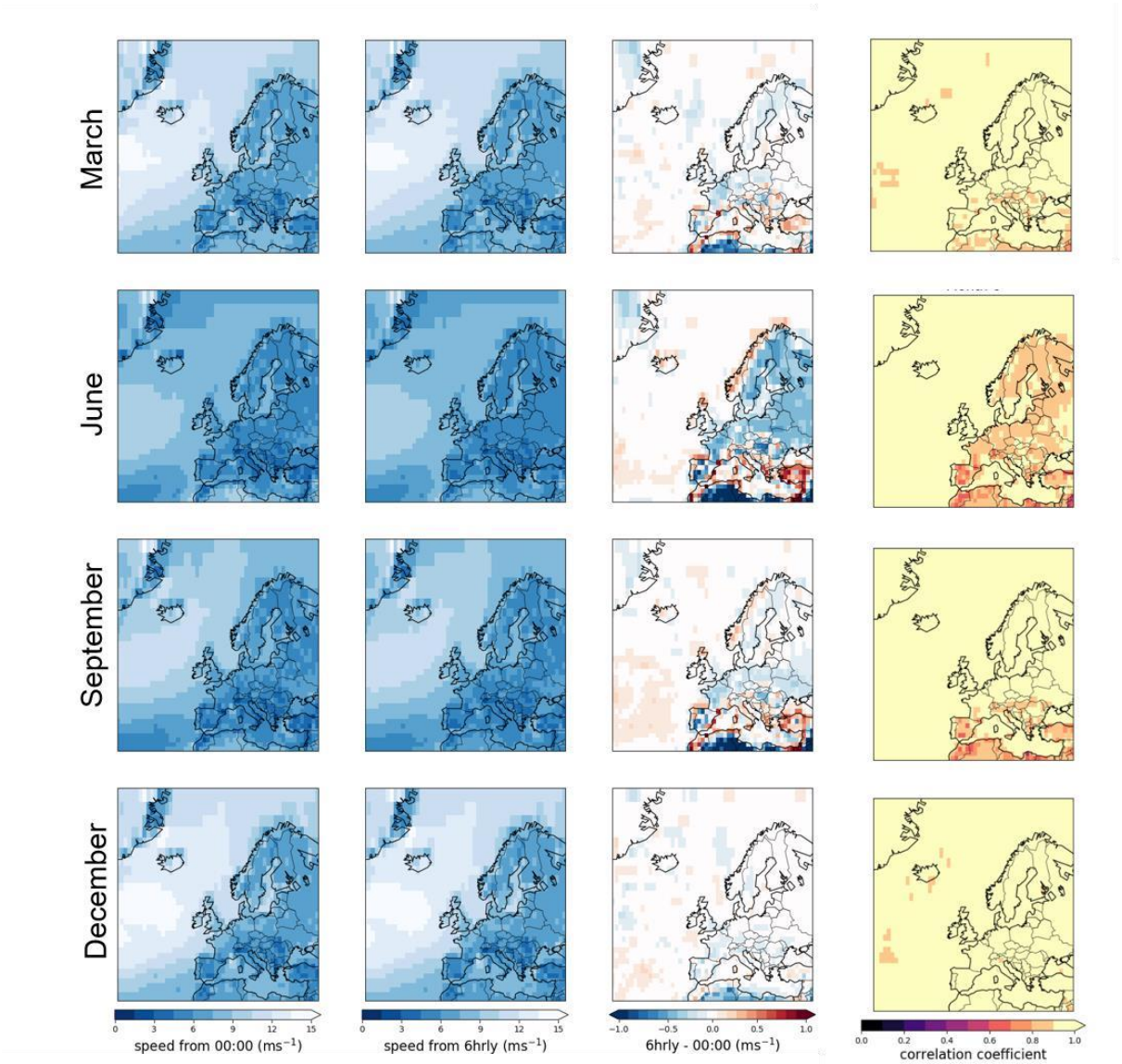


Figure A 2. Bias corrected ERA5 100m wind speed climatology's for weekly-mean wind speeds calculated from midnight winds (first column) six-hourly wind speeds (second column) the differences between for weekly-mean wind speeds calculated from midnight and six hourly winds (third column) and the correlation between the two sets of weekly-mean winds (fourth column). A representative month from each season is chosen.

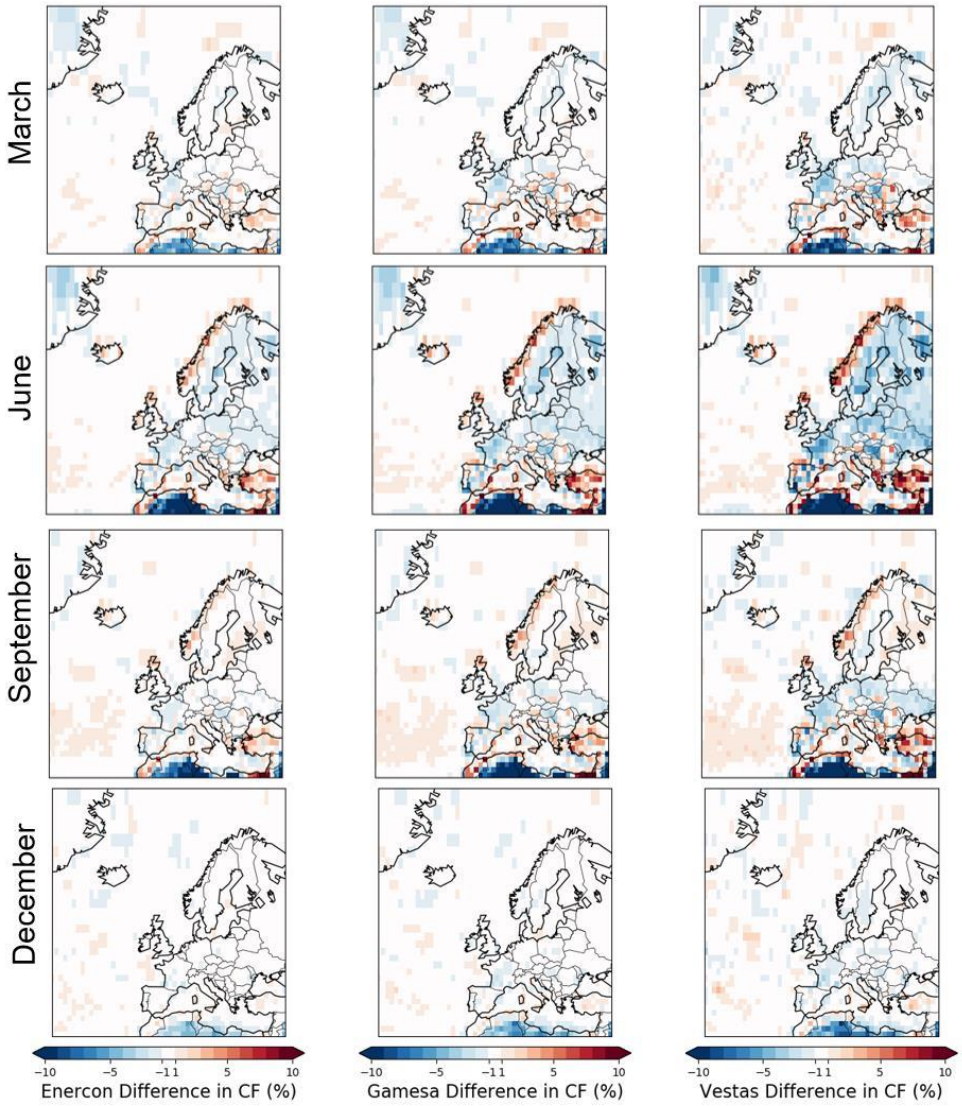


Figure A 3. Differences between weekly-mean capacity factors calculated using the bias corrected six- hourly and midnight ERA5 100m wind speeds for four representative months. Columns show the three types of wind turbine used in the DST, the Enercon (first column) Gamesa (second column) and Vestas (third column).

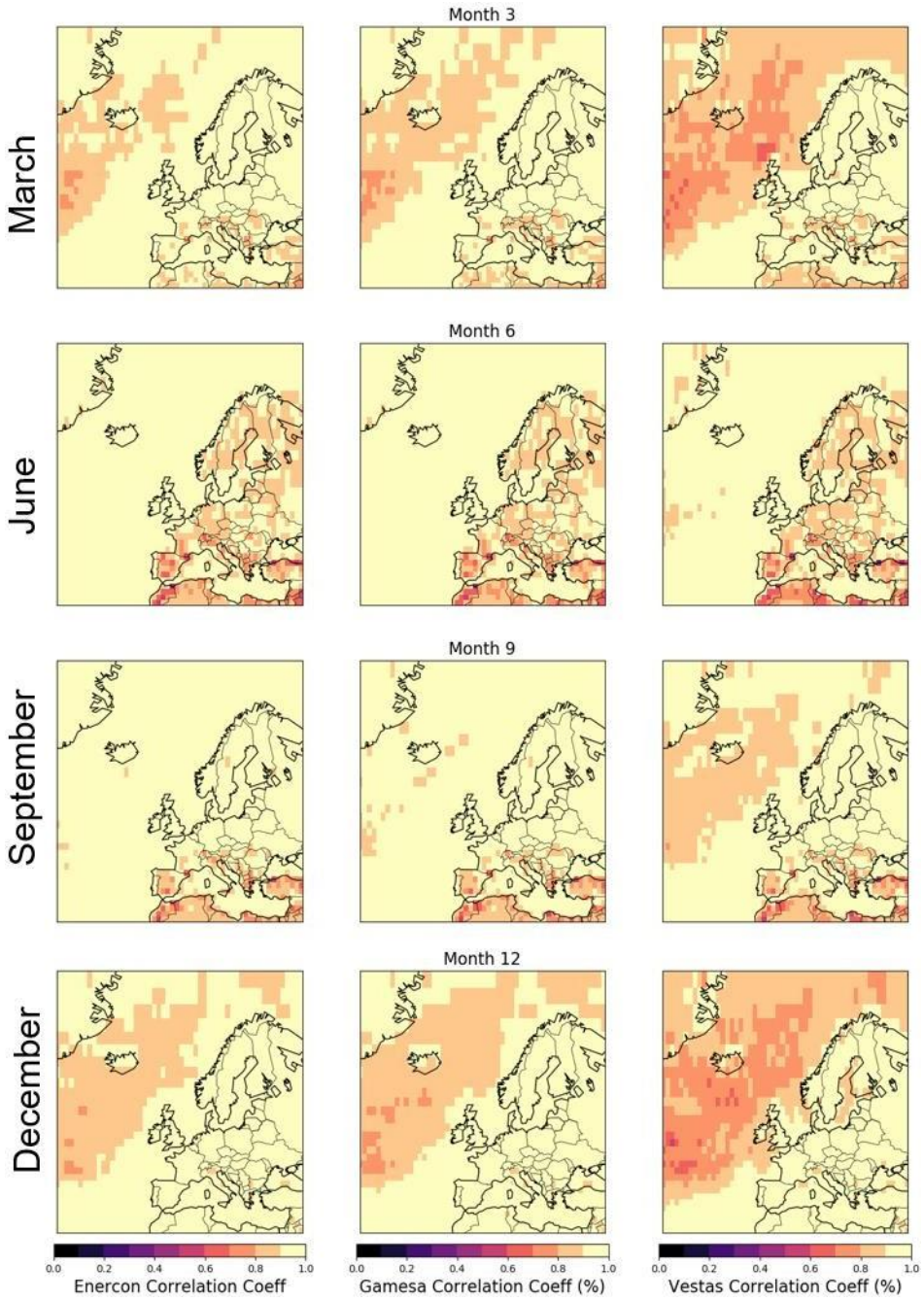


Figure A 4. Correlation between capacity factors calculated using the bias corrected ERA5 100m wind speeds for four representative months using midnight-only and six hourly data. Columns show the three types of wind turbine used in the DST, the Enercon (first column) Gamesa (second column) and Vestas (third column).

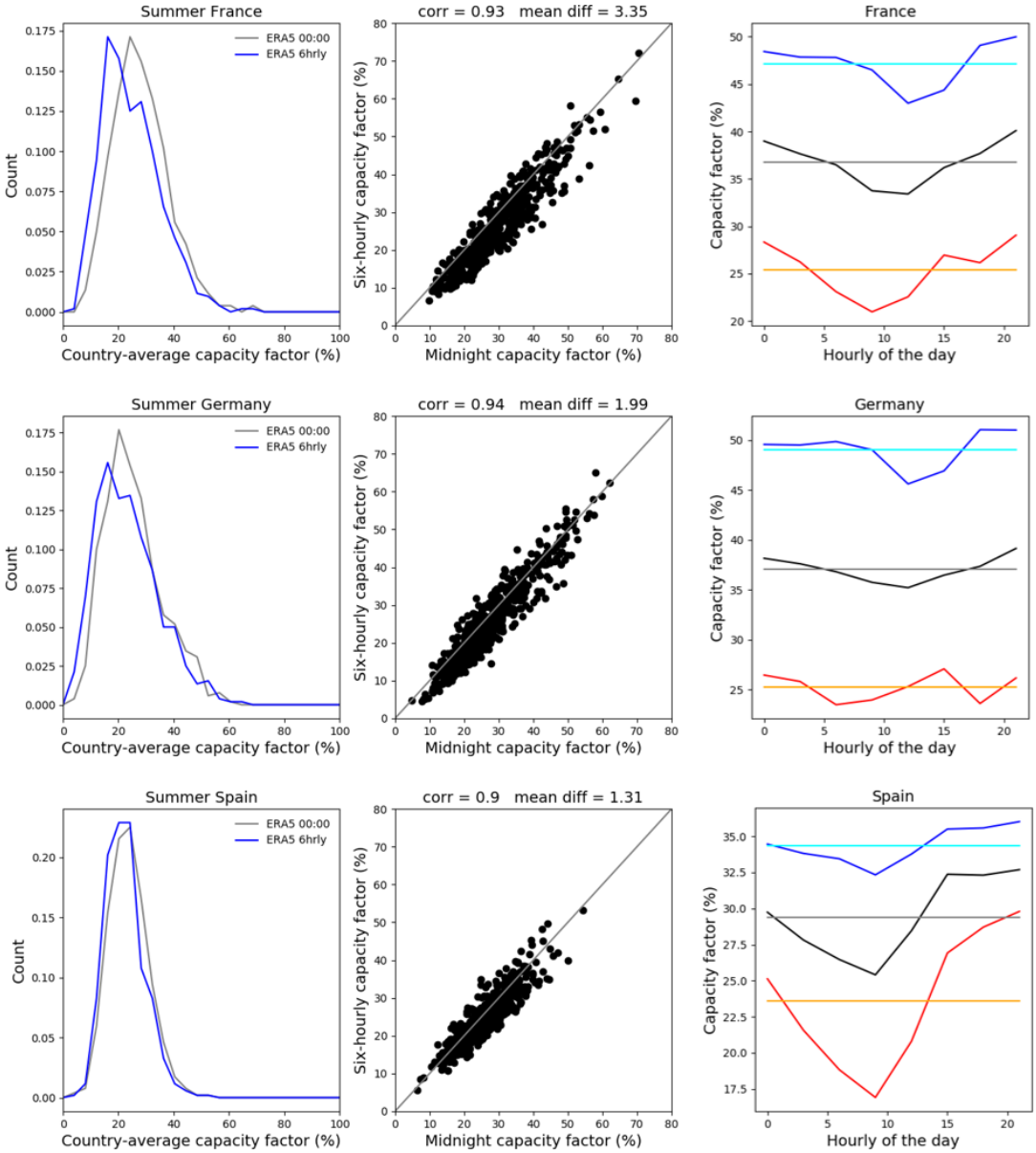


Figure A 5. Weekly-mean Country-aggregate capacity factor data for Summer (June, July and August) for the midnight only vs. six-hourly data (first column) Scatter plot showing the correlation between the two datasets (second column) A comparison of diurnal cycles of country-level capacity factor for winter (blue) summer (red) and the whole year (black), Straight lines show the daily-mean capacity factor for comparison (third column). Rows represent three case study countries, France, Germany and Spain.

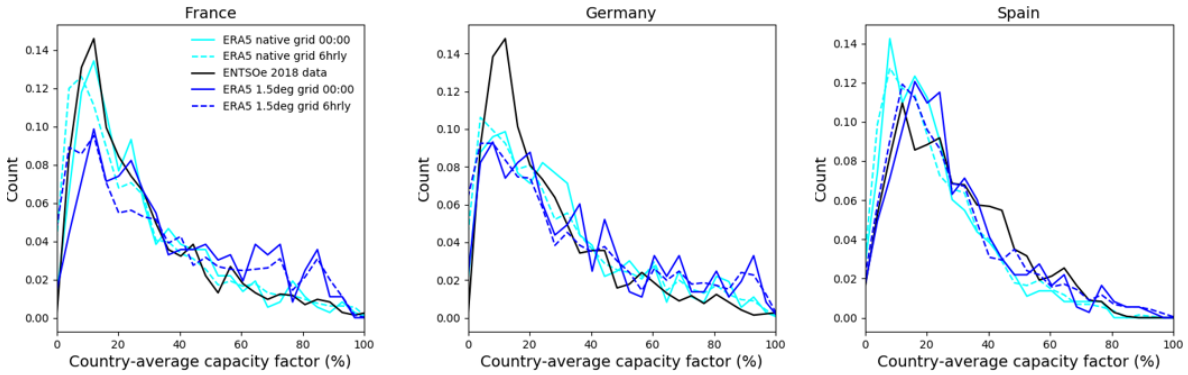


Figure A 6. Histograms of the 2018 six-hourly capacity factors from ENTSOe (black) the native resolution ERA5 wind power model (cyan) and the 1.5 degree ERA5 wind power model (blue) solid lines show six-hourly data and dotted lines show midnight-only data.

Annex 2 - Improving skill – Recommended Enhanced bias adjustment

Equations of methods used for calibration according to Torralba et al. (2017) for daily data.

Simple Correction

$$y_{Mij} = (x_{Mij} - \overline{x_M}) \frac{\sigma_{Mref}}{\sigma_{Me}} + \overline{o_M} \quad \text{(A1)}$$

where

y_{Mij}: daily adjusted forecast in month *M* for year *i* and member *j*

x_{Mij}: daily forecast in month *M* for year *i* and member *j*

$\overline{x_M}$: forecast average of month *M* of all years *i* and members *j*

$\overline{o_M}$: reanalysis average of month *M* of all years *i*

σ_{Mref}: reanalysis standard deviation for month *M*

σ_{Me}: forecast standard deviation for month *M* of all members

Calibration

$$y_{Mij} = \alpha(x_{Mij} - \overline{x_M}) + \beta(x_{Mij} - x_{Mi}) + \overline{o_M} \quad \text{(A2)}$$

Equation A2 was adapted from Torralba et al. where:

x_{Mi}: forecast average of month *M* of all members of year i

$\alpha = abs(\rho_M) \frac{\sigma_{Mref}}{\sigma_{Mem}}$ and $\beta = \sqrt{(1-\rho_M^2)} \frac{\sigma_{Mref}}{\sigma_{Me}}$

σ_{Mem}**:** forecast ensemble mean standard deviation for month *M*

σ_M: Correlation for month M of the forecast ensemble mean with the reanalysis

Please note that except y_{Mij} and x_{Mi} which are daily values**,** all coefficients of Equation A1 are monthly averages or statistics and applied on daily forecast values.

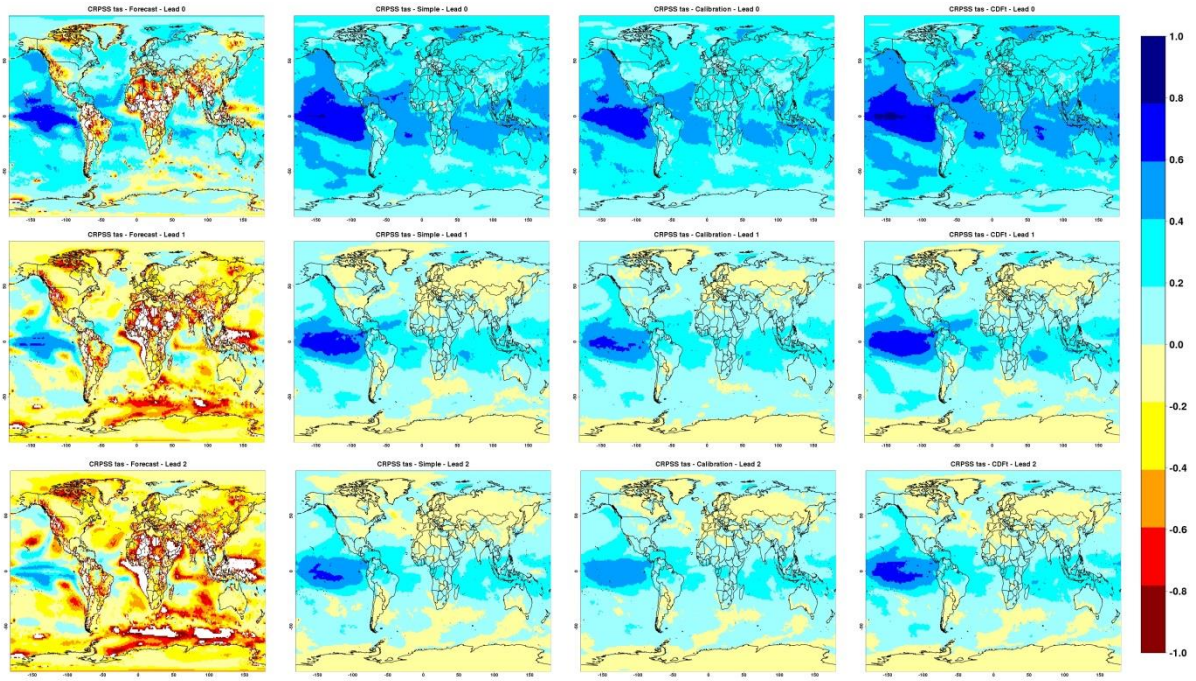


Figure A 7. Annual CRPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted monthly with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).

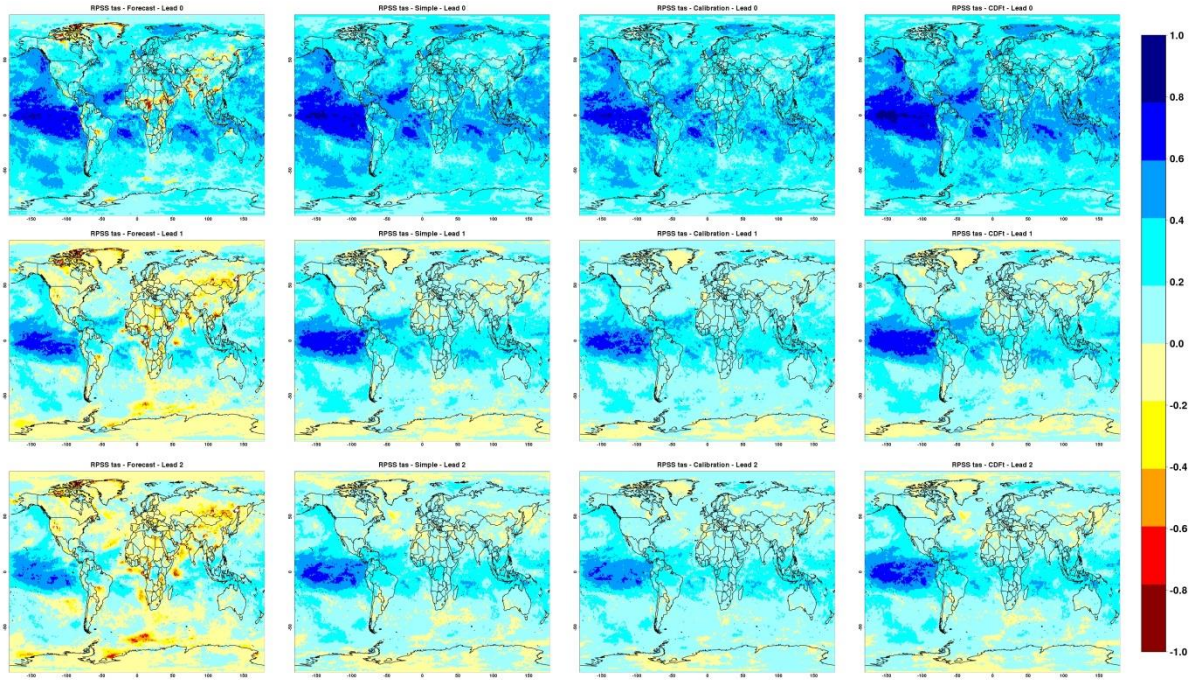


Figure A 8. Annual RPSS score of monthly averages for 2m temperature for three lead times (Top: Lead 0, middle Lead 1, bottom Lead 2) of the uncorrected forecast and the forecast adjusted monthly with the three different methods (from left to right: uncorrected forecast, simple bias adjustment, calibration and CDFt).

		Lead 0			Lead 1			Lead 2		
Domain	Forecast	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**
Europe	Forecast	-0.01	0.67	0.22	-0.27	0.27	0.00	-0.32	0.21	-0.04
	Simple	0.24	0.69	0.27	0.00	0.27	0.05	-0.02	0.22	0.03
	Calibration	0.23	0.69	0.25	0.02	0.32	0.07	0.00	0.26	0.05
	CDFt	0.27	0.69	0.28	0.02	0.27	0.06	0.00	0.22	0.04
Africa	Forecast	-0.49	0.73	0.17	-0.67	0.47	-0.01	-0.73	0.40	-0.06
	Simple	0.30	0.74	0.33	0.10	0.48	0.14	0.06	0.41	0.11
	Calibration	0.28	0.75	0.31	0.11	0.51	0.15	0.08	0.45	0.12
	CDFt	0.33	0.75	0.35	0.13	0.48	0.16	0.08	0.42	0.12
East Asia	Forecast	-0.17	0.75	0.21	-0.46	0.42	0.01	-0.53	0.37	-0.04
	Simple	0.31	0.76	0.31	0.08	0.43	0.12	0.05	0.38	0.09
	Calibration	0.28	0.76	0.29	0.09	0.46	0.13	0.07	0.40	0.11
	CDFt	0.34	0.76	0.33	0.10	0.43	0.14	0.07	0.38	0.11
North America	Forecast	0.03	0.73	0.26	-0.25	0.43	0.03	-0.33	0.37	-0.01
	Simple	0.32	0.74	0.34	0.09	0.43	0.12	0.05	0.37	0.09
	Calibration	0.29	0.75	0.32	0.10	0.46	0.13	0.07	0.40	0.11
	CDFt	0.35	0.75	0.37	0.11	0.44	0.13	0.07	0.37	0.10

Table A 1. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for 2m temperature daily averages for the first three lead times of the uncorrected forecast and the forecast adjusted monthly with the three different methods spatially averaged over Europe, Africa, East-Asia and North America.

		Lead 0			Lead 1			Lead 2		
Domain	Forecast	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**	**CRPSS**	**ACC**	**RPSS**
Europe	Forecast	-49	0.45	0.03	-157	0.05	-0.22	-275	0.03	-0.36
	Simple	0.02	0.46	0.07	-0.14	0.05	-0.10	-0.12	0.03	-0.08
	Calibration	0.05	0.49	0.11	-0.06	0.07	-0.02	-0.06	0.04	-0.01
	CDFt	0.05	0.46	0.10	-0.12	0.06	-0.09	-0.10	0.03	-0.06
Africa	Forecast	-55	0.45	-0.07	-172	0.13	-0.58	-293	0.10	-1.08
	Simple	0.03	0.47	0.00	-0.14	0.16	-0.17	-0.12	0.14	-0.15
	Calibration	0.04	0.50	-0.04	-0.07	0.20	-0.12	-0.07	0.17	-0.10
	CDFt	0.09	0.47	0.07	-0.09	0.16	-0.11	-0.08	0.13	-0.10
East Asia	Forecast	-51	0.47	0.04	-165	0.15	-0.35	-287	0.13	-0.66
	Simple	0.06	0.49	0.11	-0.13	0.16	-0.10	-0.12	0.14	-0.09
	Calibration	0.07	0.51	0.11	-0.05	0.19	-0.01	-0.05	0.17	0.00
	CDFt	0.09	0.49	0.15	-0.10	0.16	-0.07	-0.09	0.14	-0.05
North America	Forecast	-47	0.43	0.03	-157	0.08	-0.25	-277	0.08	-0.46
	Simple	0.02	0.44	0.07	-0.13	0.09	-0.09	-0.11	0.09	-0.07
	Calibration	0.05	0.47	0.09	-0.06	0.12	-0.02	-0.05	0.12	-0.01
	CDFt	0.06	0.44	0.11	-0.10	0.09	-0.06	-0.08	0.09	-0.05

Table A 2. Annual CRPSS, Anomaly Correlation Coefficient ACC, and RPSS for precipitation averages for the first three lead times of the uncorrected forecast and the forecast adjusted monthly with the three different methods spatially averaged over Europe, Africa, East-Asia and North America.

Annex 3 - Improving skill – Towards a multi-model approach – A hydrological investigation

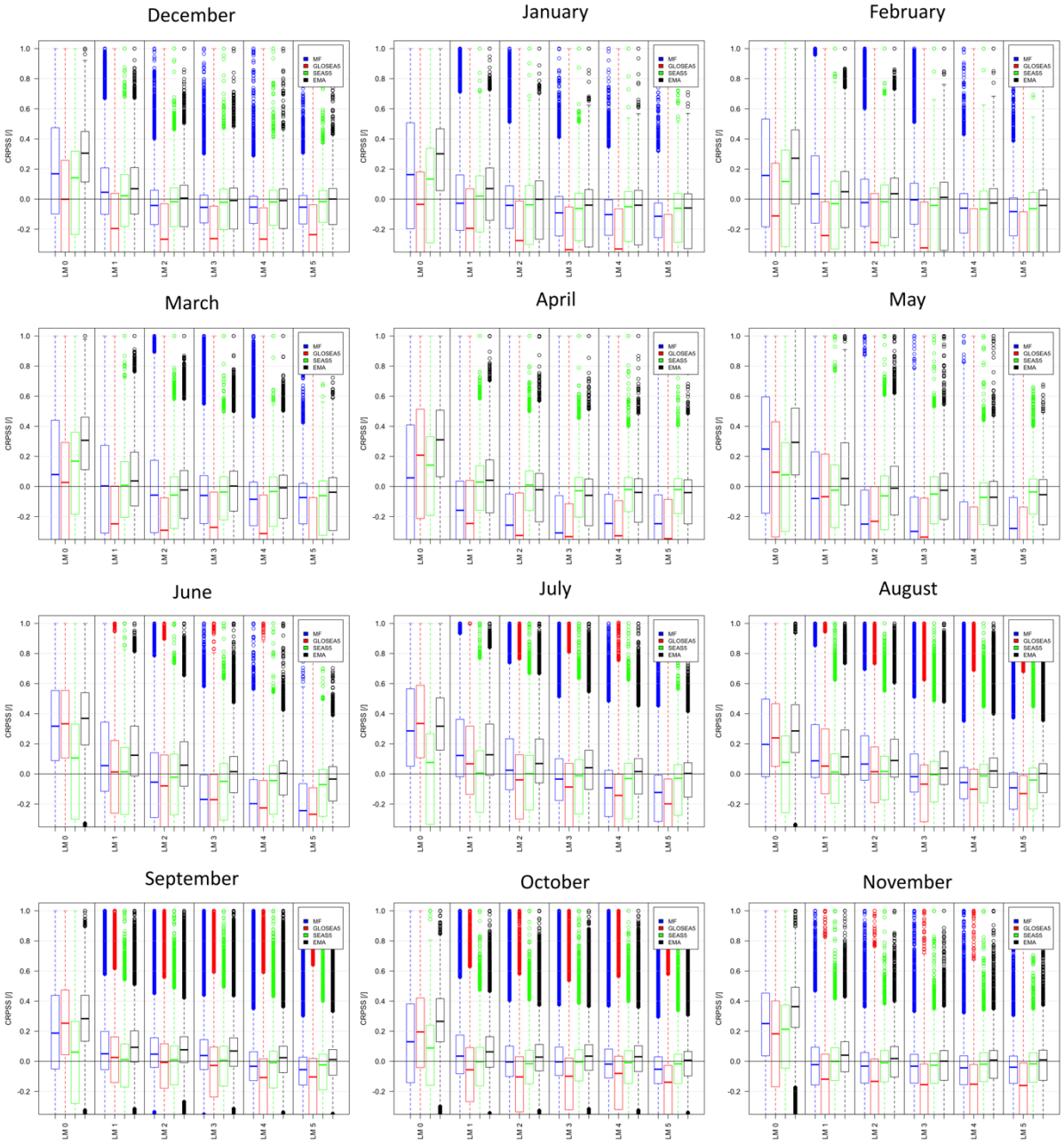


Figure A 9. Distribution of CRPSS values of individual systems (MF, GLOSEA5 and SEAS5) and multi-model (EMA) seasonal streamflow forecasts for all European sub-basins and months (Jan.-Dec.).