A new approach to extended-range multi-model forecasting: sequential learning algorithms

Gonzalez, P. L. M. ORCID: https://orcid.org/0000-0003-0154-0087, Brayshaw, D. J. ORCID: https://orcid.org/0000-0002-3927-4362 and Ziel, F. (2021) A new approach to extended-range multi-model forecasting: sequential learning algorithms. Quarterly Journal of the Royal Meteorological Society, 147 (741). pp. 4269-4289. ISSN 1477-870X doi: 10.1002/QJ.4177 Available at https://centaur.reading.ac.uk/100502/

Received: 11 January 2021 | Revised: 14 September 2021 | Accepted: 23 September 2021

DOI: 10.1002/qj.4177

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

A new approach to extended-range multimodel forecasting: Sequential learning algorithms

Paula L. M. Gonzalez[1,2,3] | **David J. Brayshaw**[1,2] | **Florian Ziel**[4]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]NCAS–Climate, University of Reading, Reading, UK

[3]International Research Institute for Climate and Society, The Earth Institute, Columbia University, Palisades, New York

[4]Universität Duisburg–Essen, Essen, Germany

Correspondence
P. L. M. Gonzalez, Department of Meteorology, University of Reading, Reading, RG6 6BG, UK.
Email: p.gonzalez@reading.ac.uk

Funding information
European Commission Horizon 2020 S2S4E: Sub-seasonal to seasonal forecasting for energy, Grant/Award Number: 776787

Abstract

Multimodel combinations are a well-established methodology in weather and climate prediction and their benefits have been widely discussed in the literature. Typical approaches involve combining the output of different numerical weather prediction (NWP) models using constant weighting factors, either uniformly distributed or determined through a prior skill assessment. This strategy, however, can lead to suboptimal levels of skill, as the performance of NWP models can vary with time (e.g., seasonally varying skill, changes in the forecasting system). Moreover, standard combination methods are not designed to incorporate predictions derived from sources other than NWP systems (e.g., climatological or time-series forecasts). New algorithms developed within the machine learning community provide the opportunity for "online prediction" (also referred to as "sequential learning"). These methods consider a set of weighted predictors or "experts" to produce subsequent predictions in which the combination or "mixture" is updated at each step to optimize a loss or skill function. The predictors are highly flexible and can combine both NWP and statistically derived forecasts transparently. A set of these online prediction methods is tested and compared with standard multimodel combination techniques to assess their usefulness. The methods are general and can be applied to any model-derived predictand. A set of weather-sensitive European country-aggregate energy variables (electricity demand and wind power) is selected for demonstration purposes. Results show that these innovative methods exhibit significant skill improvements (i.e., between 5 and 15% improvement in the probabilistic skill) with respect to standard multimodel combination techniques for lead weeks up to 5. The incorporation of statistically derived predictors (based on historical climate data) alongside NWP forecasts is also shown to contribute significant skill improvements in many cases.

KEYWORDS

expert advice, extended range, machine learning, multimodel, online prediction, s2s forecasting, sequential learning, subseasonal

1 | INTRODUCTION

Subseasonal-to-seasonal predictions (hereafter s2s), which range from a few weeks to a few months ahead, fill in the gap between weather prediction and seasonal forecasting. In the past, these forecasting ranges have received little attention, due to the limited predictability of weather forecasts beyond 10 days and the need for longer aggregation periods in order for the effects of boundary conditions to become relevant (e.g., Kirtman *et al.*, 2014; Robertson *et al.*, 2015; Vitart *et al.*, 2015). These forecasting horizons are, nonetheless, of significant relevance for a diverse range of applications (e.g., health, agriculture, water management, humanitarian efforts, energy, etc.) and have been identified as a key research area for climate services development (e.g., Vaughan *et al.*, 2016; White *et al.*, 2017; Vitart and Robertson, 2019). Many different subseasonal hindcasts and operational forecasts are now available, though these differ in important methodological aspects such as launch dates, ensemble sizes, and calibration strategies. A central challenge for all climate service sectors is therefore to derive and maximize the predictive skill available across multiple forecast systems.

The benefits of multimodel combinations in climate forecasting have previously been introduced and described for different temporal scales (e.g. Krishnamurti *et al.*, 2000; DelSole, 2007; Tebaldi and Knutti, 2007; Min *et al.*, 2009; Sansom *et al.*, 2013; Siegert and Stephenson, 2019). Most typical combination methodologies involve weighting strategies that assign each model a constant factor, either uniformly or based on an ensemble member skill assessment. Given that the skill of the models can vary at different timescales, and for multiple reasons (e.g., seasonally varying skill, or changes in the forecasting system), the fact that these weights remain constant is a methodological limitation, since it does not allow the resulting combination to adapt to these changes in skill. Other commonly used multimodel combination methodologies such as Bayesian model averaging (BMA: e.g., Raftery *et al.*, 2005) rely on strict assumptions about the probability distributions of the forecasts to obtain a posterior density.

Dynamical prediction systems are known to have limitations on these timescales, given that the lead times are too long to retain much memory of the initial conditions, but too short to be controlled by the boundary conditions (e.g., Vitart *et al.*, 2012). In this context, Cohen *et al.* (2019) propose that investing in machine-learning techniques for subseasonal forecasting may offer skill advantages over both existing numerical prediction systems and traditional statistical techniques using fixed training periods (e.g., canonical correlation analysis, CCA) alone. A large body of research exists on statistical postprocessing of climate prediction and future projections, including applications of machine-learning techniques (e.g. Monteleoni *et al.*, 2011; DelSole *et al.*, 2015) . Furthermore, statistical postprocessing of model output has often led to enhanced understanding of the physical processes linked to predictability and has therefore led to subsequent improvements to dynamical forecasting systems, and their value can therefore extend beyond forecast skill enhancement (e.g., Doblas-Reyes *et al.*, 2013).

Within the realm of machine learning, a family of algorithms has been developed to perform "online prediction with expert advice" (Cesa-Bianchi and Lugosi, 2006), also known as sequential learning or sequential aggregation rules. These methods consider a set of predictors which, after initially being weighted uniformly, produce subsequent predictions in which the combination or "mixture" is updated continually over time to optimize a loss or skill function. No assumptions are made about the probabilistic properties of the predictors or the forecasts, and the algorithms learn dynamically how the sets are linked. These online prediction methods have several potential advantages for their use in climate prediction.

- The predictor combination is updated in every forecast step, allowing the system to adjust under certain conditions (e.g., time-varying skill or new forecast system releases) to preserve or maximize skill.
- A different combination of predictors can be obtained for different quantiles of the predictand distribution to produce a robust system that maximizes skill for the full forecast probability distribution.
- The risk of including incompetent or counterproductive predictors is minimized by the system adjusting over time to assign them minimal weights.

Sequential learning algorithms have had very little application in weather and climate multimodel prediction. Mallet *et al.* (2009) first applied a sequential aggregation algorithm to combine multimodel predictions of hourly and daily near-surface ozone concentrations into a deterministic prediction. In this case, they applied a method called exponentiated gradient (Cesa-Bianchi, 1999) and showed that the sequential learning algorithm produced predictions that outperformed the best model and had similar skill to the optimal linear combination. Subsequently, Mallet (2010) enhanced the methodology further through coupling it with a data assimilation scheme, so that the sequential learning algorithm forecasted the model's analyses instead of the observations. In subsequent years, similar implementations were used to forecast other pollutants (e.g., Auder *et al.*, 2016), ocean waves (e.g., O'Donncha *et al.* 2018; 2019), and several nonmeteorological variables (e.g. Devaine *et al.*, 2013;

TABLE 1 Main features of the s2s reforecast systems considered in the study

Model	Range	Period	Frequency	Size
ECMWF ENS-extended	0–46 days	Past 20 years	2/week	11 members
NCEP CFSv2	0–42 days	1999–2010	Daily	4 members
Lagged NCEP CFSv2	0–39 days	1999–2010	2/week	12 members

Hawelka *et al.*, 2017; Amat *et al.*, 2018). Strobach and Bel (2015, 2016) first applied a similar method in climate forecasting, in the context of decadal prediction of atmospheric variables such as 2-m temperature. They also found that the exponentiated gradient algorithm was able to outperform the individual models and other standard benchmarks such as linear combinations and the climatology. Recently, Strobach and Bel (2020) applied the same exponentiated gradient learning method in the context of future climate projections using the CMIP5 ensemble (Taylor *et al.*, 2012). The authors used a historical period of reanalysis data as a learning set. In addition, they obtained an estimation of multimodel uncertainty by looking at model weight fluctuations during the learning period. Those weights and uncertainties were then used to combine future climate projections, which proved to be less uncertain than the original multimodel ensemble.

A set of limitations can be identified in the application of sequential learning algorithms mentioned above. All but Strobach and Bel (2020) focus only on producing a deterministic or point-based mean forecast. They also rely on a fixed learning rate (instead of a time-varying "adaptive" learning rate), which determines the speed at which the algorithm can adjust the combination weights (e.g., Wintenberger, 2017). A fixed learning rate prevents these methods from capturing optimally the temporal variations in the skill of the different predictors considered. Finally, these examples have drawn the predictors in a very simple way (i.e., individual forecasts) and have not explored the potential to use a more complex set of predictors to maximize skill. Sequential learning algorithms allow for improvements in all of these aspects and this work presents results that demonstrate the method's advantages.

The present work tests a set of sequential learning algorithms, with and without time-varying or "adaptive" learning rates, to combine s2s multimodel predictions and to compare the results with more traditional multimodel combination techniques. Two specific algorithms (Bernstein Online Aggregation (BOA) and the exponentiated gradient algorithm (EGA)) were tested considering the following research questions.

- To what extent do these sequential learning algorithms outperform more common "constant-weight" multimodel combinations?
- Can the skill of multimodel forecast combinations be improved through the incorporation of reanalysis-based predictors?

Within the context of the EU-Horizon 2020 S2S4E project (Subseasonal-to-seasonal Forecasting for Energy[1]), these novel methods are applied to forecasts of national-average electricity demand and wind-power generation across a set of European countries, though the methodology can readily be applied to other s2s forecast properties.

2 | DATA AND METHODOLOGY

2.1 | Dynamical subseasonal predictions

Extended-range numerical weather predictions (NWP) and their corresponding reforecasts were compiled by the S2S prediction project[2] (Vitart *et al.*, 2017).

Hindcast ensembles from the European Centre for Medium-range Weather Forecasting (ECMWF: ECMWF ENS-ER[3]) and National Centers for Environmental Prediction (NCEP: CFS version 2, Saha *et al.*, 2014) modeling centers were obtained from the ECMWF S2S data portal[4] and are described in Table 1.

In the case of ECMWF ENS-ER, the system has an on-the-fly reforecast methodology that generates hindcasts for the 20 years prior to the forecasts, with 11 ensemble members launched biweekly on Mondays and Thursdays. All hindcasts matching forecasts initialized during the calendar year 2016 are included in the analysis, and therefore three separate model cycles had to be considered:

- CY41R1: for hindcasts from January 1–March 7;
- CY41R2: for hindcasts from March 8–November 21; and
- CY43R1: for hindcasts from November 22–December 31.

[1]https://s2s4e.eu/.
[2]http://s2sprediction.net/.
[3]https://www.ECMWF.int/en/forecasts/documentation-and-support/extended-range-forecasts.
[4]https://apps.ECMWF.int/datasets/data/s2s-realtime-instantaneous-accum-ecmf/.

In the case of NCEP, the system has a fixed reforecast period from 1999–2010 and hindcasts were launched daily, creating a four-member ensemble. Given the small size of the ensemble but the frequent launch dates, a lagged ensemble is used to obtain a dataset with a structure and size comparable to ECMWF prior to the multimodel combination: the daily starts of NCEP were subsampled biweekly to match the ECMWF starts (Mondays and Thursdays) and each of those starts was combined with the two preceding ones to generate a larger number of members (12 instead of 4).

The analysis is then performed on the common period for the two datasets, which is 1999–2010. That 12-year period gets reduced to 2000–2010, since the first year is lost to generate one-year persistence forecasts (see Section 2.4). Initial tests revealed that it takes around two years for the aggregation methodologies to achieve a quasi-equilibrium in their weights variability. Therefore, the following two years (2000–2001) are used to optimize algorithm parameters when necessary, and all skill evaluations are restricted to the nine-year period 2002–2010.

Daily mean 2-m temperatures were obtained for each of the hindcasts and their mean biases were adjusted using equivalent variables from the ERA5 reanalysis (Hersbach *et al.*, 2020). A lead-time-dependent bias correction to the mean is applied, forcing the climatology to match that of ERA5. Furthermore, a variance inflation is then implemented to adjust the hindcasts' spread to match ERA5's while preserving their correlation (Doblas-Reyes *et al.*, 2005). These adjustments are applied at each grid point by first regridding the ERA5 variables to the 1.5° model grid. A leave-one-out approach is implemented on the hindcast (e.g., the calibration for hindcast year 1999 uses the hindcast climatology of years 2000–2010). A sample time-series plot of UK electricity demand in the different datasets, together with the subperiods of the analysis, is included as Figure S1 in the Supporting Information.

2.2 | Electricity demand

To illustrate the use of online prediction on an application-relevant variable, daily grid-point 2-m temperatures from both the reanalysis and hindcasts sets are converted to country-aggregate daily electricity demand time series, using a statistical data model (Bloomfield *et al.*, 2020) that accounts for the weather-driven variability while removing time-evolving socio-economic drivers, such as weekly cycles and long-term trends. Grid-point electricity demand time series are aggregated to the country level using European countries' shapefiles (Bloomfield *et al.*, 2020) and compared with the corresponding time series derived from the ERA5 reanalysis (Hersbach *et al.*, 2020) for the same 1999–2010 period.

The daily hindcasts were then used to create weekly averages with the following criteria:

- week 1: days 1–7;
- week 2: days 8–14;
- week 3: days 15–21;
- week 4: days 22–28;
- week 5: days 29–35.

We highlight, however, that the application to electricity demand is included as a proof of concept. The method demonstrated below is very general and can be applied to any forecast variable or derived magnitude. Some additional examples can be found in the Supporting Information (see Section S5).

2.3 | Sequential learning algorithms: a qualitative description

As a first introduction to sequential learning algorithms (SLAs), we present here a conceptual description of the methodology. A more complete description of the theory of SLAs and all the relevant formulations are included in the Supporting Information (see Section S1.1). This section is meant to describe the "mechanics" of the SLAs in plain language.

Given an observable time series of interest $Y(t)$, these methods consider a set of predictors $E(n,t)$ to be combined into a subsequent prediction of $Y(t)$, for example $\widehat{Y}(t+1)$. The methods consider a linear combination of the predictors and start by assigning them equal weights. Nonetheless, in subsequent time steps (i.e., $t+2$, $t+3$, etc.), the objective of these methods is progressively to update the set of weights $W(n,t+i)$ so that the resulting prediction $\widehat{Y}(t+i)$ minimizes a certain loss function of the individual forecasts (i.e., a given measure of error of the predictions). The SLAs can be adjusted to control how quickly the combination weights are allowed to change between forecast steps (referred to as the "learning rate"), how many prior steps are considered in the determination of the updated weights (referred to as the "forgetting factor"), and whether or not there is a lower-bound constraint on the individual weights (referred to as the "fixed share", and typically used to avoid some predictors being discarded permanently).

This article presents comparative results from implementing two different SLA algorithms, one with a time-varying learning rate, Bernstein Online Aggregation

(BOA: Wintenberger, 2017), and one with a constant learning rate, the Exponentiated Gradient Algorithm (EGA: Kivinen and Warmuth, 1997). Due to the adaptive learning rates, the BOA algorithm allows for more flexibility in the updating process over time when comparing with the EGA. In each case, a set of improvements to the standard "off-the-shelf" implementations was applied by optimizing the aforementioned parameters over an initial training period. Details about the methods and these implementations can be found in the Supporting Information (see Section S2).

2.4 | Multimodel combinations and skill references

This study tests the implementation of sequential learning algorithms for the prediction of weekly hindcasts of country-aggregate electricity demand. With this purpose, a simple set of predictors was considered, split into two categories: NWP-based and reanalysis-based. It is emphasized that this predictor set could be expanded, potentially to improve the resulting output forecast (e.g., by including more NWP forecast systems, earlier launch dates, pattern-based indicators or other information sources), but this simple predictor set is sufficient for demonstration purposes.

NWP-based

- 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 (denoted with _1) and lagged NCEP (denoted with _2).
- Minimum and maximum of the ensemble distribution (**FCST_MN, FCST_MX**). For each start, the minimum and maximum values from both hindcast systems are retained.

Reanalysis-based

- Quantiles of the climatology (**Q10_CLIM, Q35_CLIM, Q50_CLIM, Q65_CLIM, Q90_CLIM**). Obtained from the 1.5° ERA5 climatology for each calendar day using a leave-one-out approach on the years.
- Persistence (**PERS**). Weekly persistence forecast based on ERA5, calculated using the seven days prior to each hindcast start date.
- Last-year persistence (**PERS_1yr**). Weekly persistence forecast based on ERA5's demand for the same week of the previous calendar year.

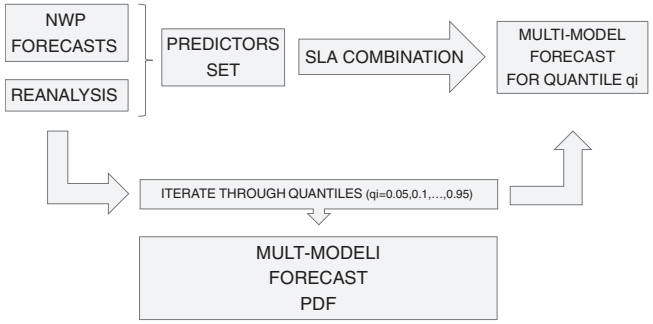

FIGURE 1 Diagram summarizing the application of sequential learning algorithms to combine NWP forecasts and reanalysis into a multimodel probabilistic forecast

- Seasonal minimum and maximum (**SEAS_MN, SEAS_MX**). For each hindcast week, the minimum and maximum values in the ERA5 climatology are obtained using a leave-one-out approach on the hindcast year.

The multimodel forecasting method considered in this study is described by the diagram in Figure 1. The predictors described above were combined, fully or partially, using different combinations or aggregation methodologies to obtain a multimodel forecast for a given quantile of the distribution, as described in Section S1.1. By iterating over a set of quantiles ("Qgrid"), the method results in a multimodel probabilistic forecast. As a simple illustration, consider the EGA algorithm applied to combine the NWP predictors (Q10–Q90, FCST_MIN and FCST_MAX) from two systems NCEP and ECMWF. For a given forecast quantile (e.g., the 5th percentile), a new prediction is derived by obtaining combination weights for each of the input predictors. The process is then repeated for the 10th and 15th–95th percentiles (i.e., a separate set of weights is calculated in each case) such that a probabilistic forecast is produced. Given the limited sizes of the reforecast systems considered here, a quantile spacing of 0.05 (5%) was considered. Additionally, a different aggregation rule was created for each hindcast lead (weeks 1–5).

Each mixture was applied using all predictors (denoted as BOA, EGA) and the NWP-based predictors only (denoted as BOA_NWP and EGA_NWP) in order to assess the presence of any added value from including reanalysis-based information in the forecasts. The EGA method was implemented to benchmark the results against prior uses of sequential learning algorithms in climate prediction (e.g., Strobach and Bel 2015; 2016; 2020), but it has been considered here with several improvements with respect to its prior uses: EGA is trained for each q_i in "Qgrid" using a 2-year training period and optimizing a fixed learning rate across quantiles over this period.

Two further innovations were tested and introduced which proved to increase skill: the use of an optimized "forgetting factor", which controls the relevance that older predictions have in the determination of the updated weights, and the use of an optimized "fixed share", which assigns a minimal baseline weight to all the predictors to stop them from being discarded permanently from predictions. These strategies are further described in the Supporting Information (see Section S3). A summary table for all the algorithms is included in the section (see Table S1).

The multimodel combinations were then benchmarked against a set of references described below.

- NWP-based equal-weights combination (**EW_NWP**). For each q_i in "Qgrid", a quantile forecast is obtained from each system (ECMWF and NCEP) and they are then averaged with equal weights (in this case, $w = 0.5$).
- Ensemble model output statistic calibration (**EMOS**). A Gaussian EMOS ensemble calibration, also known as nonhomogeneous Gaussian regression (Gneiting *et al.*, 2005), was trained using the probabilistic predictions of the NWP systems (**ECMWF, NCEP**) from the first TWO forecast years (128 time steps) and then used to generate calibrated probabilistic predictions for the evaluation period and for each q_i in "Qgrid". This scheme was considered as an example of ensemble statistical postprocessing and calibration techniques used in probabilistic forecasting.
- Climatology (**CLIM**). For each q_i in "Qgrid", the 1.5° ERA5 was used to create a climatology for each calendar day using a leave-one-out approach on the years.
- Individual NWP systems (**ECMWF, NCEP**). For each q in "Qgrid", a quantile forecast is obtained from the ensemble members of each system.
- Oracles (**O_NWP_conv, O_NWP_lin**). Using the full evaluation period, the oracles are built as the optimal multiple linear regression combinations of all the NWP-based predictors, 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. Because these oracles require the full knowledge of the entire evaluation period, they provide an upper boundary-estimate for the skill of a fixed-weight performance-based multimodel combination. In an operational setting, "oracle" combinations are not possible, because they require knowledge of the full period, and analogous linear regressions would need to be trained using a previously observed independent period.

It is worth mentioning that, even though persistence is another typical benchmark in weather and climate forecasting, it is not used as such here, because it is part of the predictors set.

3 | RESULTS

The skill of the sequential learning algorithms (SLAs) is first examined in the context of deterministic or point-based skill, as in prior applications of the algorithms for climate prediction (e.g., Strobach and Bel 2015; 2016; 2020). As a subsequent step, the skill improvements of the methods are assessed in a probabilistic context, considering all the quantiles in "Qgrid".

3.1 | Deterministic skill: application to United Kingdom demand

3.1.1 | Forecast skill for the ensemble median (50th quantile)

To illustrate the aggregation methods and their evaluation, we first consider the case of United Kingdom (UK) electricity demand. As an initial step, the skill of the SLA combinations is evaluated using the 50th quantile (Q50) as a measure of the average behavior of the ensemble. As described in Section 2.3, the methods are optimized in terms of a loss function, which in this case was the pinball or quantile loss (see Section S1.1), which is calculated for each prediction time. The average pinball loss for Q50 is equivalent to the mean absolute error (MAE) and is therefore a metric of the deterministic skill of the resulting system. Figure 2 presents these time-averaged losses for the proposed aggregation rules and the benchmark forecasts.

Figure 2a shows the full time-average losses, sorted according to lead week 3, and shows that all the SLAs (i.e., all the versions of the EGA and BOA methods considered here) are more skillful than the equal-weights combination (EW_NWP), the EMOS calibration, the climatology (CLIM), and the individual NWP systems (ECMWF, NCEP) for that lead time. Figure 2b presents the relative skill improvements of the predictions relative to the EW_NWP combination, calculated as the ratio between the pinball losses for each week. Three other aspects of the results are highlighted. Firstly, all the SLAs present an increase in skill with respect to EW_NWP for every lead time. This, added to the fact that the oracle combinations (O_NWP_conv and O_NWP_lin, the optimal constant-weight combinations) are less skillful than some of the combinations, indicates that there are potential benefits of SLAs compared with constant-weight combinations and more sophisticated calibrations such as

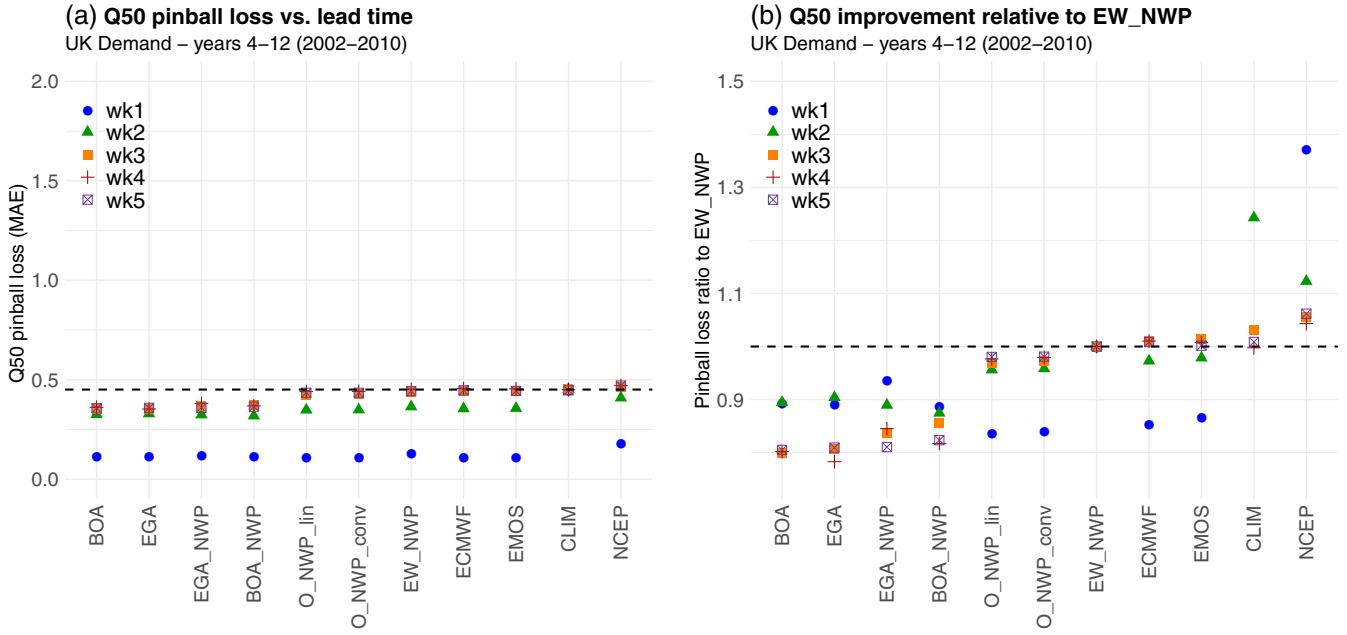

FIGURE 2 (a) UK demand Q50 average pinball loss associated with the aggregation rules and the reference forecasts. The average losses for weeks 1–5 are presented as different symbols. The items on the *x*-axis are sorted from smaller to larger based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for climatology, as a reference. (b) Relative improvement in skill expressed as relative losses with respect to the equal-weights combination (EW_NWP) for each corresponding week

EMOS. The SLAs implemented here resulted in a significant reduction in the pinball loss in weeks 1–5 with respect to all the other predictions (ranging between around 9% and 22%). Finally, these plots also suggest that some benefit arises from including reanalysis-based predictors in the combinations, as the full versions typically have lower losses than the NWP-only SLAs.

3.1.2 | Average weights for the 50th quantile

To explore the composition of the aggregation methods and the differences in their performances, Figure 3 presents the time-average weights for different weekly leads and all versions of the BOA and EGA combinations. It shows that, for week 1, the higher weights are assigned to the quantiles of the ECMWF ensemble, though not necessarily centered at Q50. This suggests that the aggregation is correcting residual biases in the shape of the forecasts distribution, which were not addressed by the initial mean correction and variance inflation (see Section 2.1). The NCEP quantiles also present a noticeable contribution to the combinations, though smaller than 10%. As lead time increases (Figure 3b and c, corresponding to weeks 3 and 5), the combinations that include reanalysis-based predictors (black bars) shift weight to the quantiles of the climatology, whereas, for the aggregations that only include NWP-based predictors (BOA_NWP, EGA_NWP), larger weights are assigned to the quantiles of NCEP than in shorter leads. Also, the FCST_MX and FCST_MN gain some relevance.

Differences between the BOA and EGA algorithms are noticeable in the relative distributions of the weights, such as along the quantiles of the ECMWF forecasts, but not so much in the evolution with lead time.

It is relevant to point out that, even though this section has discussed the features of the average weights, the fact that they are time-varying is a significant property of the SLAs. As an example, the temporal evolution of a set of weights is discussed in the Supporting Information (see Section S4).

3.2 | Probabilistic skill: pinball loss as a function of quantile

The prior evaluation of the aggregation rules was focused on the center of the forecast distribution and associated with the deterministic skill. Nonetheless, these aggregation techniques have the potential to generate improvements in skill throughout the forecast distribution. To assess this capability objectively, one can study the behavior of the average pinball loss as a function of the distribution quantile. Figure 4 presents such an analysis for weeks (a) 2 and (b) 5 for UK electricity demand, where each line corresponds to the pinball loss

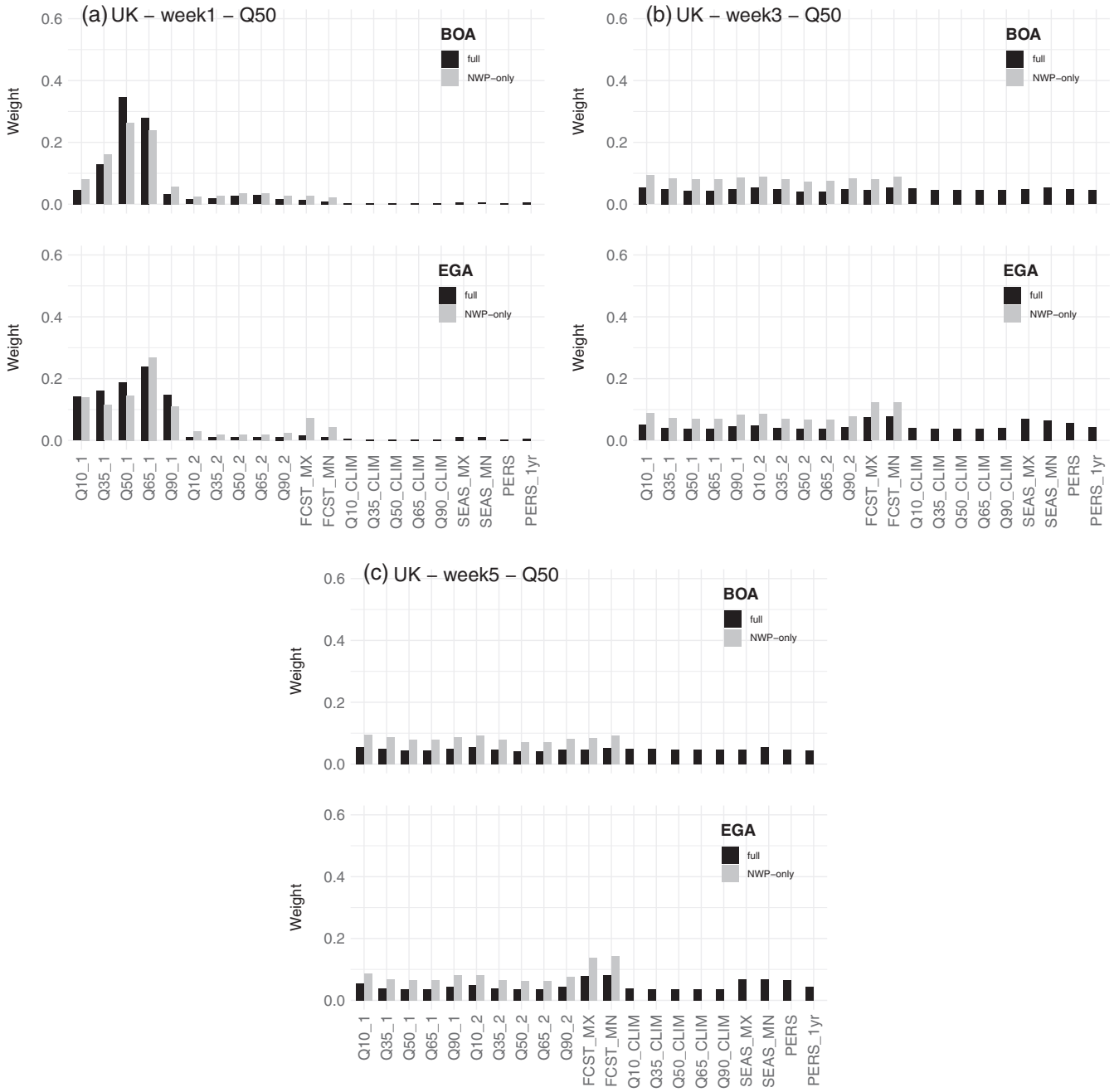

FIGURE 3 Average weights obtained for the UK demand Q50 aggregation rules of the BOA and EGA algorithms over weeks (a) 1, (b) 3, and (c) 5. The black bars correspond to the combinations that consider all the predictors and the gray bars to the NWP-only aggregations. Predictors identified with _1 correspond to the ECMWF system and those marked with _2 to the NCEP system

of a model combination or a reference forecast relative to EW_NWP.

For week 2 (Figure 4a), the climatology and NCEP are clearly outperformed by EW_NWP for every quantile. ECMWF, however, outperforms EW_NWP everywhere but on the lower quantiles. All the other combinations have smaller losses compared with the uniform combination for most quantiles, but the differences between them are relatively small. It can be seen that, for week 2, the four SLA combinations have clear improvements in the center of the distribution, whereas the oracle combinations show smaller losses in the tails. It is worth recalling that the oracles are not a fair benchmark, since they represent an upper bound to the skill of constant-weights linear combinations (i.e., they were trained using the full period). In general, the EGA algorithms are less stable across the distribution and can become less skillful for particular quantiles.

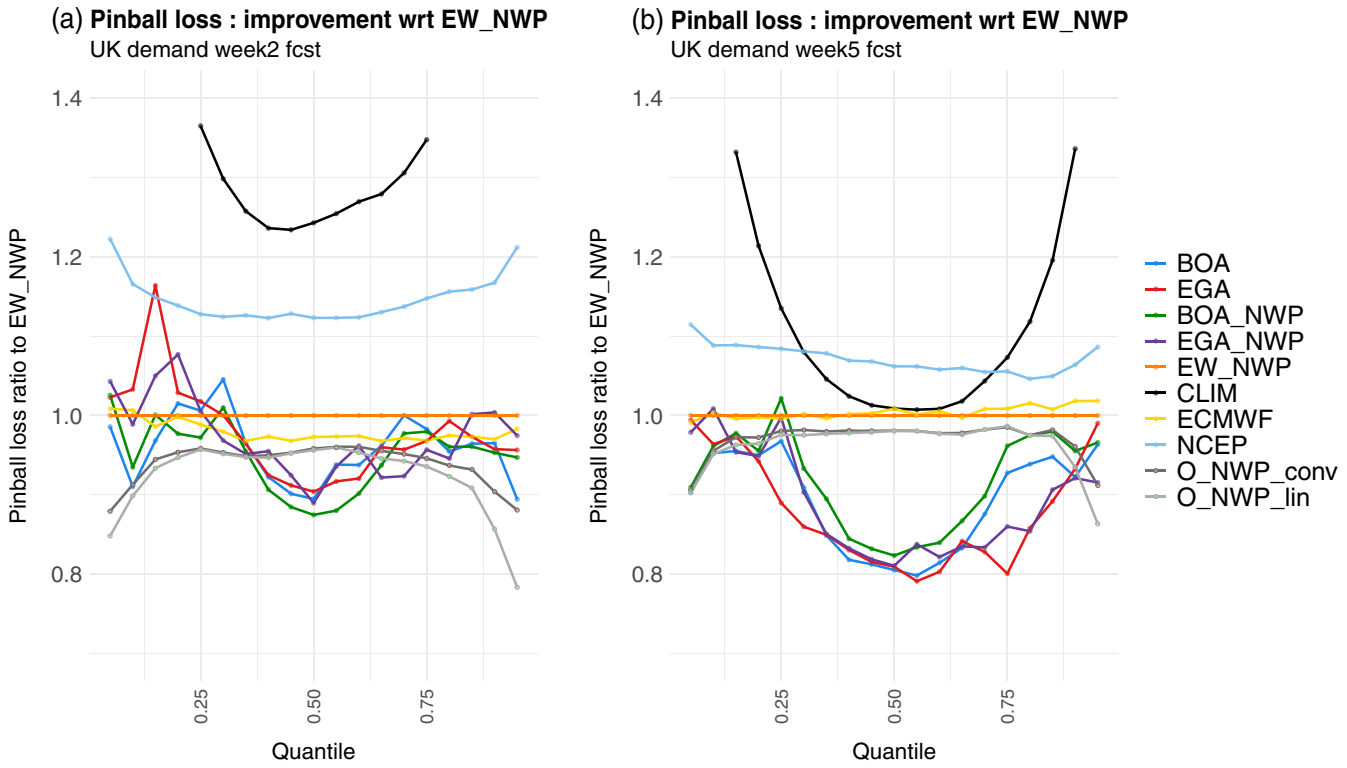

FIGURE 4 Average pinball loss as a function of the quantile for UK demand forecasted as (a) week 2 and (b) week 5. Losses are presented as the relative difference with respect to the EW_NWP loss corresponding to each week. The oracles have been marked in gray and hollow dots to highlight them, as they represent an upper bound to the skill of constant-weights linear combinations

In the case of week 5 (Figure 4b), the differences in losses between the forecasts and references become larger. NCEP and the climatology are both outperformed by EW_NWP for every quantile, whereas ECMWF remains very close in skill to EW_NWP throughout the distribution. The BOA and EGA combinations now exhibit the largest skill improvements for most quantiles, only being outperformed in the extreme tails by the oracles. It is important to note that these tails are subject to larger errors due to the limited size of the sample analyzed here, and therefore these disagreements might not be robust. For this lead week, there is a suggestion that the combinations that include reanalysis-based predictors are more skillful than their NWP-only counterparts for a large part of the distribution, in particular in the case of the BOA algorithm. This suggests that incorporating reanalysis-based information in the online process might result in an increase in skill for longer lead times.

3.2.1 | Quantile-mean pinball loss

Rather than focusing on any particular quantile of the distribution, one might want to evaluate the overall skill of the resulting combinations. To address that, the average of the losses along Qgrid can be computed. It can be shown that, provided Qgrid is fine enough, this quantile mean pinball loss is a good approximation for the continuous ranked probability score (Taieb *et al.*, 2016). The continuous ranked probability score (CRPS: Matheson and Winkler, 1976) measures the distance between the observed and the predicted cumulative distribution functions (e.g., Hersbach, 2000) and it is widely used in the evaluation of probabilistic forecasting skill.

Figure 5a presents the quantile-mean losses (CRPS estimates) for all the combinations and references. In agreement with the previous section, the four SLA combinations show the smallest quantile mean losses for lead weeks 3–5 (Figure 5a). Additionally, the combinations that include reanalysis-based predictors beat their _NWP counterparts in every case. All the sequential aggregation algorithms show improvements with respect to EW_NWP and the climatology (Figure 5b) for weeks 1–5. The relative improvements in the CRPS estimate of the SLAs for weeks 3–5 range between around 7% and 16%, whereas for week 2 the skill improvements remain close to 5%. For week 1, improvements can be quite high with respect to the EW_NWP, but the SLA combinations do not outperform ECMWF.

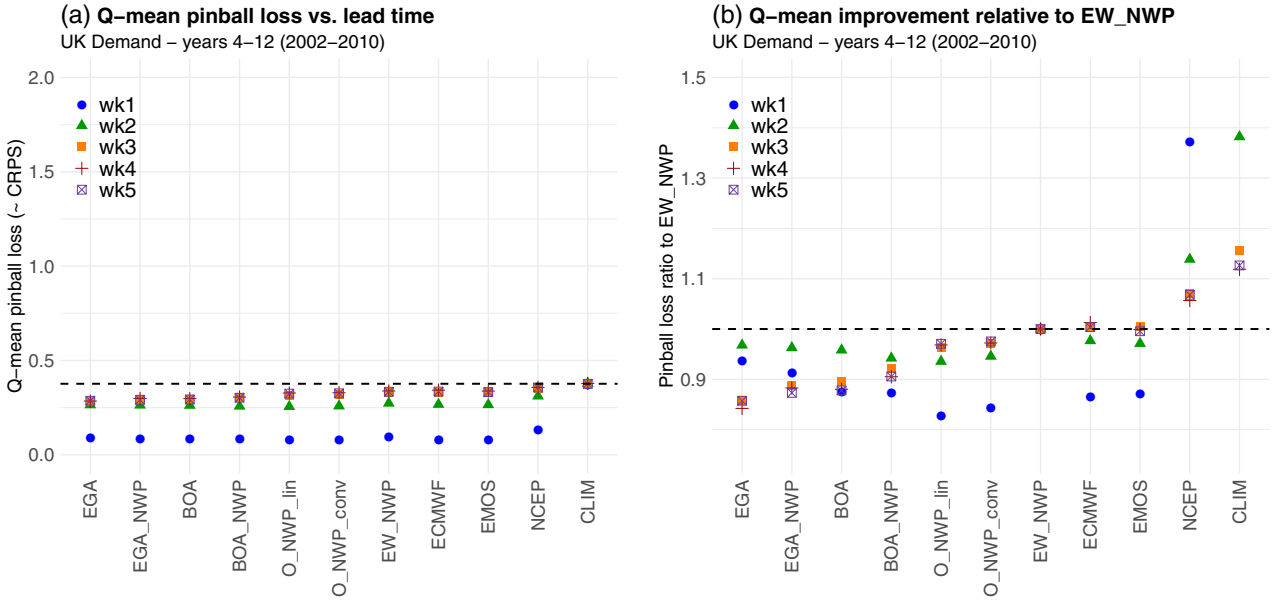

FIGURE 5 (a) UK demand quantile-mean average pinball loss (CRPS estimate) associated with the aggregation rules and the reference forecasts. The average losses for weeks 1–5 are presented as different symbols. The items on the *x*-axis are sorted from smaller to larger loss based on the week 3 results (orange squares). The dashed horizontal black line shows the value of the week 1 loss for the climatology, as a reference. (b) Relative improvement in skill expressed as relative losses with respect to the EW_NWP combination, for each corresponding week

It is also important to remark that most combinations are more skillful than the oracles for weeks 3–5. This is of relevance, because the linear combination of forecasts (such as through ridge regression) is another standard multimodel forecasting technique used in weather and climate prediction. The oracles, however, provide an upper limit for the skill of constant-weights combinations, since they were optimized using the complete evaluation period, when in practice a fixed independent training period would have been used to obtain the weights. The SLAs, however, can be more skillful than the oracles by allowing the combination weights to vary with time. In a similar way, the SLA combinations outperform the EMOS probabilistic forecast calibration for weeks 3–5.

3.3 | Statistical significance of skill improvements

The increases in skill presented above, though moderate in some cases, are quite robust. The statistical significance of the skill enhancements can be assessed by using a Diebold–Mariano (DM) test (Diebold and Mariano, 1995), which compares the predictive accuracy of two forecasts using the time series of associated quantile-mean losses.

Results from the DM tests applied to UK demand forecasts are presented in Table 2 and reveal that all the combinations resulted in forecasts that are significantly more skillful than EW_NWP for weeks 1–5. Also, the increased skill of EGA with respect to BOA algorithms appears to be significant only for longer leads. Similarly, the advantages of including reanalysis-based predictors become significant for long leads (3–5 in most cases).

The assessment above might be affected by the multiple testing or multiple comparisons effect (e.g., Benjamini and Hochberg, 1995), but the application of corrections to the significance levels to make the tests more stringent is not universally recommended (e.g., Rothman, 1990; Rubin, 2021). In this context, we have decided to complement the analysis above with the determination of Model Confidence Sets (MCSs: Hansen *et al.*, 2011), selected objectively for each metric and lead week. This method considers the full set of models and the null hypothesis of equal predictive power (or equally, that no inferior model is present in the set). Given a significance level, the procedure sequentially removes the worst-performing model until the null hypothesis can no longer be rejected. The remaining models are therefore the model confidence set for that confidence level. We implemented this method using the `R`-package `estMCS`[5] and considering the quantile-mean pinball losses, as in the case of the DM test, for each week individually. The optimal block length for bootstrapping was selected using the method

[5]https://rdrr.io/github/nielsaka/modelconf/man/estMCS.html.

TABLE 2 Results from a Diebold–Mariano significance test applied to pairs of UK electricity demand forecasts

Comparison	Week 1	Week 2	Week 3	Week 4	Week 5
BOA versus EW_NWP	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
EGA versus EW_NWP	**<0.001**	*0.042*	**<0.001**	**<0.001**	**<0.001**
BOA_NWP versus EW_NWP	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
EGA_NWP versus EW_NWP	**<0.001**	**<0.001**	*0.028*	**<0.001**	**<0.001**
EGA versus BOA	1.000	0.781	**0.001**	**0.002**	*0.037*
BOA versus BOA_NWP	0.760	0.982	**<0.001**	**<0.001**	**<0.001**
EGA versus EGA_NWP	0.992	0.694	**<0.001**	**<0.001**	*0.022*

Note: The values correspond to the *p*-value of the test. Bold numbers indicate the accuracy improvement is significant at the 99% (*p*-value ≤ .01) level and italics indicate levels higher than 95% (*p*-value ≤ .05).

TABLE 3 Model confidence sets (MCSs) obtained from UK electricity demand forecasts for each lead week

Forecast lead	UK demand MCS
Week 1	O_NWP_lin
Week 2	BOA, EGA, BOA_NWP, EGA_NWP, O_NWP_lin, O_NWP_conv
Week 3	EGA
Week 4	BOA, EGA
Week 5	BOA, EGA, EGA_NWP

Note: The values were obtained considering a 99% confidence threshold.

proposed by Patton *et al.* (2009) and implemented using the `R`-package `np`.[6]

The resulting MCSs for UK demand forecasts are presented in Table 3. The resulting sets imply that, at the 99% confidence level, no inferior model is contained. The results of this stringent significance test are in line with the assessment presented above, and show that the SLA combinations outperform all the other methods for lead weeks 3–5. For lead week 2, the SLA combinations show similar skill levels to the oracles, but they still outperform the equal-weight combination and the individual NWP models. For lead weeks 3 and 4, it is also shown that combinations that include reanalysis-based predictors outperform the NWP-only ones.

4 | DISCUSSION AND CONCLUSIONS

An innovative set of multimodel aggregation techniques was introduced through application to subseasonal-to-seasonal UK weekly electricity demand forecasting. The results included here have shown very promising results, with a significant skill improvement, particularly at long lead times (i.e., beyond week 3).

Even though some initial benefits of the sequential learning algorithms were obtained when analyzing the deterministic skill of the forecasts (Q50), the full extent of the skill improvements is observed when accounting for the complete quantile distribution. The two multimodel aggregation algorithms tested here, EGA and BOA, showed significant skill improvements with respect to the climatology, standard multimodel methods, and the individual best NWP system (ECMWF) for weeks 2–5. Overall, the optimized BOA and EGA combinations resulted in average improvements ranging around between 7 and 16% in the quantile-mean pinball losses (an estimation for the CRPS) for weeks 3, 4, and 5.

In the results presented here, some additional skill was obtained when the combinations included reanalysis-based predictors. A case study included in the Supporting Information illustrates how the algorithms are able to "learn" from the performance of the predictors through a season and adjust the weights accordingly to minimize losses (see Section S4). These results suggest that, when reanalysis-based predictors were included, this adjustment resulted in better predictions.

These methods were also tested on other large countries such as Germany, France, and Spain, for both electricity demand and wind power, and the main results are included in the Supporting Information (see Section S5). The overall conclusions remain robust, though the relative comparisons with the forecast benchmarks depend on the region and method.

We therefore conclude that the application of these novel multimodel combination techniques is very promising for s2s prediction. Given that the use of these techniques has little computational cost, more complex

[6]https://cran.r-project.org/web/packages/np/np.pdf.

implementations would also be feasible, including, for example, the following: prior launches of the same forecasting system as predictors; different types of predictors, such as pattern- or regime-based forecasts; or a truly seamless approach to s2s forecasting by incorporating prior seasonal forecasts. Moreover, the "online" nature of these techniques makes them particularly suitable for operational practice: weights are automatically updated with every forecast cycle (rather than relying on some *a priori* calibration process using a "fixed" sample). Updates and changes in observing systems or NWP models can therefore be incorporated seamlessly without any additional effort.

It is, however, important to note that these SLAs should not be treated as a "black box", and careful implementation and performance verification is required. In particular, while it is clear that the SLAs can always be implemented (i.e., standard versions are available through open-source statistical packages), the process through which skill enhancements are achieved remains unclear. On the one hand, the fact that the SLAs show the largest skill improvements for longer lead times suggests that they act, in part, as an additional bias correction process (i.e., removing residual forecast biases that remained after the initial calibration of the raw NWP model output). On the other hand, the analysis of case studies showing "shocks" in the weight evolution suggests that the SLAs benefit from reanalysis-based predictors that carry memory-like information about how the recent past is behaving with respect to climatology and other predictors. This study introduced the use of SLAs to s2s prediction as a proof of concept, but more research is necessary to identify objectively the sources of the significant skill improvements associated with these methods, and to understand how these might interact with a more complex set of predictors and with additional skill assessment metrics. Additionally, it would be important to assess the skill of these novel methods against more complex multimodel benchmarks that are being applied in s2s prediction (e.g., Wanders and Wood, 2016; Specq and Batté, 2020).

ACKNOWLEDGMENTS

The work of P.L.M.G. and D.J.B. has been funded by the Horizon 2020 EC project number 776787 ("S2S4E: Sub-seasonal to seasonal forecasting for energy"). P.L.M.G. was also supported by the National Centre for Atmospheric Science. The authors thank Dr Hannah Bloomfield for allowing the use of her models and datasets and Dr David Livings for supporting preprocessing of the hindcast and reanalysis data used in this study. We also thank Matteo De Felice for his insightful comments on an initial version of the manuscript. The authors also thank two anonymous reviewers for their feedback, which led to a clearer presentation of the results of this work.

AUTHOR CONTRIBUTIONS

Paula L. M. Gonzalez: conceptualization; data curation; formal analysis; investigation; methodology; resources; validation; visualization; writing – original draft; writing – review and editing. **David J. Brayshaw:** conceptualization; funding acquisition; investigation; project administration; supervision; writing – original draft; writing – review and editing. **Florian Ziel:** conceptualization; formal analysis; investigation; methodology; resources; validation; visualization; writing – original draft; writing – review and editing.

ORCID

Paula L. M. Gonzalez https://orcid.org/0000-0003-0154-0087

REFERENCES

Amat, C., Michalski, T. and Stoltz, G. (2018) Fundamentals and exchange rate forecastability with simple machine learning methods. *Journal of International Money and Finance*, 88, 1–24.

Auder, B., Bobbia, M., Poggi, J.-M. and Portier, B. (2016) Sequential aggregation of heterogeneous experts for pm10 forecasting. *Atmospheric Pollution Research*, 7, 1101–1109.

Büeler, D., Beerli, R., Wernli, H. and Grams, C.M. (2020) Stratospheric influence on ECMWF sub-seasonal forecast skill for energy-industry-relevant surface weather in European countries. *Quarterly Journal of the Royal Meteorological Society*, 146, 3675–3694. https://doi.org/10.1002/qj.3866.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300.

Bloomfield, H.C., Brayshaw, D.J. and Charlton-Perez, A.J. (2020) Characterizing the winter meteorological drivers of the European electricity system using targeted circulation types. *Meteorological Applications*, 27, e1858. https://doi.org/10.1002/met.1858.

Cesa-Bianchi, N. (1999) Analysis of two gradient-based algorithms for on-line regression. *Journal of Computer and System Sciences*, 59, 392–411.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G. and Stoltz, G. (2012) Mirror descent meets fixed share (and feels no regret). Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q. eds. *Advances in Neural Information Processing Systems*, 25, Red Hook, NY: Curran Associates, Inc., 980–988.

Cesa-Bianchi, N. and Lugosi, G. (2006) *Prediction, Learning, and Games*. Cambridge, UK: Cambridge University Press.

Cohen, J., Coumou, D., Hwang, J., Mackey, L., Orenstein, P., Totz, S. and Tziperman, E. (2019) S2s reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *WIREs Climate Change*, 10, e00567. https://doi.org/10.1002/wcc.567.

DelSole, T. (2007) A Bayesian framework for multimodel regression. *Journal of Climate*, 20, 2810–2826.

DelSole, T., Monteleoni, C., McQuade, S., Tippett, M., Pegion, K. and Shukla, J. (2015) *Tracking seasonal prediction models*. Dy, J. G., Emile-Geay, J., Lakshmanan, V., & Liu, Y. eds. Proceedings of the Fifth International Workshop on Climate Informatics.

Devaine, M., Gaillard, P., Goude, Y. and Stoltz, G. (2013) Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, 90, 231–260. https://doi.org/10.1007/s10994-012-5314-7.

Diebold, F.X. and Mariano, R.S. (1995) Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13, 253–263. https://doi.org/10.1080/07350015.1995.10524599.

Doblas-Reyes, F.J., García-Serrano, J., Lienert, F., Biescas, A.P. and Rodrigues, L.R. (2013) Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4, 245–268.

Doblas-Reyes, F.J., Hagedorn, R. and Palmer, T. (2005) The rationale behind the success of multimodel ensembles in seasonal forecasting–II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, 57, 234–252.

Gaillard, P. and Goude, Y. (2016) Software package described here: https://cran.r-project.org/web/packages/opera/opera.pdf

Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118.

Hansen, P.R., Lunde, A. and Nason, J.M. (2011) The model confidence set. *Econometrica*, 79, 453–497. https://doi.org/10.3982/ECTA5771.

Hawelka, B., Sitko, I., Kazakopoulos, P. and Beinat, E. (2017) Collective prediction of individual mobility traces for users with short data history. *PLoS One*, 12, 1–14. https://doi.org/10.1371/journal.pone.0170907.

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15, 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hong, T. and Fan, S. (2016) Probabilistic electric load forecasting: a tutorial review. *International Journal of Forecasting*, 32, 914–938.

Kirtman, B.P., Min, D., Infanti, J.M., Kinter, I., J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P. and Becker, E. (2014) The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, 95, 585–601.

Kivinen, J. and Warmuth, M.K. (1997) Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132, 1–63.

Krishnamurti, T.N., Kishtawal, C.M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S. and Surendran, S. (2000) Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13, 4196–4216. https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Lynch, K.J., Brayshaw, D.J. and Charlton-Perez, A. (2014) Verification of European subseasonal wind speed forecasts. *Monthly Weather Review*, 142, 2978–2990. https://doi.org/10.1175/MWR-D-13-00341.1.

Mallet, V. (2010) Ensemble forecast of analyses: coupling data assimilation and sequential aggregation. *Journal of Geophysical Research: Atmospheres*, 115, D24303.

Mallet, V., Stoltz, G. and Mauricette, B. (2009) Ozone ensemble forecast with machine learning algorithms. *Journal of Geophysical Research: Atmospheres*, 114, D05307.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096.

Min, Y.-M., Kryjov, V.N. and Park, C.-K. (2009) A probabilistic multimodel ensemble approach to seasonal prediction. *Weather and Forecasting*, 24, 812–828. https://doi.org/10.1175/2008WAF2222140.1.

Monteleoni, C., Schmidt, G.A., Saroha, S. and Asplund, E. (2011) Tracking climate models. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4, 372–392. https://doi.org/10.1002/sam.10126.

O'Donncha, F., Zhang, Y., Chen, B. and James, S.C. (2018) An integrated framework that combines machine learning and numerical models to improve wave-condition forecasts. *Journal of Marine Systems*, 186, 29–36.

O'Donncha, F., Zhang, Y., Chen, B. and James, S.C. (2019) Ensemble model aggregation using a computationally lightweight machine-learning model to forecast ocean waves. *Journal of Marine Systems*, 199, 103206.

Patton, A., Politis, D.N. and White, H. (2009) Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White. *Econometric Reviews*, 28, 372–375. https://doi.org/10.1080/07474930802459016.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133, 1155–1174. https://doi.org/10.1175/MWR2906.1.

Robertson, A.W., Kumar, A., Peña, M. and Vitart, F. (2015) Improving and promoting subseasonal to seasonal prediction. *Bulletin of the American Meteorological Society*, 96, ES49–ES53. https://doi.org/10.1175/BAMS-D-14-00139.1.

Rothman, K.J. (1990) No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43–46. https://www.jstor.org/stable/pdf/20065622.pdf

Rubin, M. (2021) When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese*, 1–32.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., Ek, M., Meng, J., Yang, R., Peña Mendez, M., van den Dool, H., Zhang, Q., Wang, W., Chen, M., and Becker, E. (2014) The NCEP climate forecast system version 2. *Journal of Climate*, 27, 2185–2208.

Sansom, P.G., Stephenson, D.B., Ferro, C.A., Zappa, G. and Shaffrey, L. (2013) Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments. *Journal of Climate*, 26, 4017–4037.

Siegert, S. and Stephenson, D.B. (2019). Forecast recalibration and multimodel combination. In A.W. Robertson and F. Vitart

(Eds.), *Sub-Seasonal to Seasonal Prediction*, Chapter 15. Elsevier, pp. 321–336.

Specq, D. and Batté, L. (2020) Improving subseasonal precipitation forecasts through a statistical-dynamical approach: application to the southwest tropical Pacific. *Climate Dynamics*, 55, 1913–1927.

Strobach, E. and Bel, G. (2015) Improvement of climate predictions and reduction of their uncertainties using learning algorithms. *Atmospheric Chemistry and Physics*, 15, 8631–8641.

Strobach, E. and Bel, G. (2016) Decadal climate predictions using sequential learning algorithms. *Journal of Climate*, 29, 3787–3809. https://doi.org/10.1175/JCLI-D-15-0648.1.

Strobach, E. and Bel, G. (2020) Learning algorithms allow for improved reliability and accuracy of global mean surface temperature projections. *Nature Communications*, 11(451). https://doi.org/10.1038/s41467-020-14342-9.

Taieb, S.B., Huser, R., Hyndman, R.J. and Genton, M.G. (2016) Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. *IEEE Transactions on Smart Grid*, 7, 2448–2455.

Taylor, K.E., Stouffer, R.J. and Meehl, G.A. (2012) An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1.

Tebaldi, C. and Knutti, R. (2007) The use of the multimodel ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A*, 365, 2053–2075.

Vaughan, C., Buja, L., Kruczkiewicz, A. and Goddard, L. (2016) Identifying research priorities to advance climate services. *Climate Services*, 4, 65–74.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L. (2017) The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, 98, 163–173.

Vitart, F., Robertson, A. W. and S2S Steering Group. 2015. "Sub-Seasonal to Seasonal Prediction: Linking Weather and Climate." *Chapter 20 in Seamless Prediction of the Earth System: From Minutes to Months*. Brunet, G., Jones, S., & Ruti, P. M. Eds., WMO-No 1156, World Meteorological Organization, 385–401.

Vitart, F. and Robertson, A.W. (2019). Introduction: why sub-seasonal to seasonal prediction (s2s)? In A.W. Robertson and F. Vitart (Eds.), *Sub-Seasonal to Seasonal Prediction*, Chapter 1. Elsevier, pp. 3–15.

Vitart, F., Robertson, A.W. and Anderson, D.L. (2012) Subseasonal to seasonal prediction project: bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, 61, 23.

Wanders, N. and Wood, E.F. (2016) Improved sub-seasonal meteorological forecast skill using weighted multimodel ensemble simulations. *Environmental Research Letters*, 11, 094007.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K.V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B. and Zebiak, S.E. (2017) Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications*, 24, 315–325. https://doi.org/10.1002/met.1654.

Wintenberger, O. (2017) Optimal learning with Bernstein online aggregation. *Machine Learning*, 106, 119–141.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Gonzalez, P.L.M., Brayshaw, D.J. & Ziel, F. (2021) A new approach to extended-range multimodel forecasting: Sequential learning algorithms. *Quarterly Journal of the Royal Meteorological Society*, 1–14. Available from: https://doi.org/10.1002/qj.4177