Improving sub-seasonal forecast skill of meteorological drought: a weather pattern approach

Doug Richardson[1,2], Hayley J. Fowler[2], Chris G. Kilsby[2], Robert Neal[3], Rutger Dankers[3,4]

[1]CSIRO Oceans & Atmosphere, Hobart, Australia, 7001
[2]School of Engineering, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, United Kingdom
[3]Weather Science, Met Office, Exeter, EX1 3PB, United Kingdom
[4]Wageningen Environmental Research, Wageningen University & Research, Wageningen, 6708 PB, Netherlands

Correspondence to: Doug Richardson (doug.richardson@csiro.au)

Abstract. Dynamical model skill in forecasting extratropical precipitation is limited beyond the medium-range (around 15 days), but such models are often more skilful at predicting atmospheric variables. We explore the potential benefits of using weather pattern (WP) predictions as an intermediary step in forecasting UK precipitation and meteorological drought on sub-seasonal time scales. Mean sea-level pressure forecasts from the ECMWF ensemble prediction system (ECMWF-EPS) are post-processed into probabilistic WP predictions. Then we derive precipitation estimates and dichotomous drought event probabilities by sampling from the conditional distributions of precipitation given the WPs. We compare this model to the direct precipitation and drought forecasts from ECMWF-EPS and to a baseline Markov chain WP method. A perfect-prognosis model is also tested to illustrate the potential of WPs in forecasting. Using a range of skill diagnostics, we find that the Markov model is the least skilful, while the dynamical WP model and direct precipitation forecasts have similar accuracy independent of lead-time and season. However, drought forecasts are more reliable for the dynamical WP model. Forecast skill scores are generally modest (rarely above 0.4), although those for the perfect-prognosis model highlight the potential predictability of precipitation and drought using WPs, with certain situations yielding skill scores of almost 0.8, and drought event hit and false alarm rates of 70% and 30%, respectively.

1 Introduction

Droughts are a recurrent climatic feature in the UK. Severe events, such as those in 1975-76, 1995 and 2010-12, had significant implications for many sectors, including agriculture, water resources and the economy, as well as for ecosystems and natural habitats (Marsh, 1995; Marsh *et al.*, 2007; Rodda and Marsh, 2011; Kendon *et al.*, 2013). To mitigate the effects of drought, it is crucial that relevant sectors plan ahead, and drought forecasts have an important role in designing these strategies. Despite this, there is very little published research on UK drought prediction, and studies have predominantly focussed on hydrological drought (Wedgbrow *et al.*, 2002; Wedgbrow *et al.*, 2005; Hannaford *et al.*, 2011).

Meteorological drought is challenging to predict using dynamical ensemble prediction systems (Yoon *et al.*, 2012; Dutra *et al.*, 2013; Yuan and Wood, 2013; Mwangi *et al.*, 2014; Lavaysse *et al.*, 2015). This is primarily due to the complex processes involved in precipitation formation, making it a difficult variable to forecast beyond short lead-times (Golding, 2000; Cuo *et al.*, 2011; Smith *et al.*, 2012; Saha *et al.*, 2014). At longer lead-times, dynamical model skill in predicting atmospheric variables tends to be much higher (Saha *et al.*, 2014; Scaife *et al.*, 2014; Vitart, 2014; Baker *et al.*, 2018). This has led researchers to investigate the potential of using atmospheric forecasts as a precursor to predicting precipitation-related hazards (Lavers *et al.*, 2014; Lavers *et al.*, 2016; Baker *et al.*, 2018).

Weather pattern (WP; also called weather types, circulation patterns and circulation types) classifications are a candidate for such an application. A WP classification consists of a number of individual WPs, which are typically defined by an atmospheric variable and represent the broad-scale atmospheric circulation over a given domain (Huth *et al.*, 2008). They can be used to make general predictions of local-scale variables such as wind speed, temperature and precipitation and are a tool for reducing atmospheric variability to a few discrete states. WP classifications have mainly been studied in the context of extreme hydro-meteorological events (Hay *et al.*, 1991; Wilby, 1998; Bárdossy and Filiz, 2005; Richardson *et al.*, 2018a; Richardson *et al.*,

2018b), and as a tool for analysing historical and future changes in atmospheric circulation patterns (Hay *et al.*, 1992; Wilby, 1994; Brigode *et al.*, 2018). See Huth *et al.* (2008) for a comprehensive review of WP classifications.

Until recently, the capability of dynamical models to predict WP occurrences had been little researched. Ferranti *et al.* (2015) evaluated the forecast skill of the medium-range European Centre for Medium-Range Weather Forecasts ensemble prediction system (ECMWF-EPS) (Buizza *et al.*, 2007; Vitart *et al.*, 2008) using WPs. They objectively defined four WPs according to daily 500 hPa geopotential heights over the North Atlantic – European sector. Model forecasts of this variable for October through April between 2007 and 2012 were then assigned to the closest matching WP using the root-mean-square difference. Verification scores indicated that there was superior skill for predictions initialised during negative phases of the North Atlantic Oscillation (NAO) (Walker and Bliss, 1932). Similarly, WPs were used to evaluate the skill of the Antarctic Mesoscale Prediction System by Nigro *et al.* (2011).

To support weather forecasting in the UK in the medium- to long range, the Met Office use a WP classification, MO30, in a post-processing system named "Decider" (Neal *et al.*, 2016). Using a range of ensemble prediction systems, forecast mean sea-level pressure (MSLP) fields over Europe and the North Atlantic Ocean are assigned to the best-matching WP according to the sum-of-squared differences between the forecast MSLP anomaly and WP MSLP anomaly fields. Decider therefore produces a probabilistic prediction of WP occurrences for each day in the forecast lead-time. Decider has various operational applications: predicting the possibility of flow transporting volcanic ash originating in Iceland into UK airspace, highlighting potential periods of coastal flood risk around the British Isles (Neal *et al.*, 2018) and as an early-forecast system for fluvial flooding (Richardson *et al.*, in review).

For Japan, Vuillaume and Herath (2017) defined a set of WPs according to MSLP. These WPs were used to refine bias-correction procedures, via regression modelling, of precipitation from two global ensemble forecast systems. The authors found that improvements from the bias-correction method using WPs was strongly dependent on the WP, but overall superior to the global (non-WP) method. Relevant to this study, Lavaysse *et al.* (2018) predicted monthly meteorological drought in Europe using a WP-based method. They aggregated ECMWF-EPS daily reforecasts of WPs to predict monthly frequency anomalies of each WP. For each 1° grid cell, the predictor was chosen to be the WP that corresponded to the maximum absolute temporal correlation between the monthly WP frequency of occurrence anomaly and the monthly Standardised Precipitation Index (SPI) (McKee *et al.*, 1993). Using this relationship, the model predicted drought in a grid cell when 40% of the ECMWF-EPS ensemble members forecast a Standardised Precipitation Index (SPI; McKee *et al.*, 1993) value below -1. Compared to direct ECMWF-EPS drought forecasts, the WP-based model was more skilful in north-eastern Europe during winter, but less skilful for central and eastern Europe during spring and summer. Over the UK, the WP model appeared to be superior for north-western regions in winter, but inferior in summer, although scores for the latter were of low magnitude.

The aforementioned studies have all considered daily WPs. An example of WPs defined on the seasonal time-scale was presented by Baker *et al.* (2018). The authors analysed reforecasts of UK regional winter precipitation between the winters of 1992-93 and 2011-12 using GloSea5, which has little raw skill in forecasting this variable (MacLachlan *et al.*, 2015). GloSea5 has, however, been shown to skilfully forecast the winter NAO (Scaife *et al.*, 2014). Baker *et al.* (2018) exploited this by constructing two winter MSLP indices over Europe and the North Atlantic, and reforecasts of these indices were derived from the raw MSLP fields. A simple regression model then related these indices to regional precipitation and produced more skilful forecasts than the raw model output.

In this study, we shall explore the potential for utilising a WP classification (specifically MO30) in UK meteorological drought prediction. We shall predict WPs using two models, ECMWF-EPS and a Markov chain, from which precipitation and drought forecasts will be derived. These models will be compared to direct precipitation and drought forecasts from ECMWF-EPS. We also run an idealised, perfect prognosis model that uses WP observations rather than forecasts as an 'upper benchmark' to

assess the upper limit of the usefulness of the WP classification. Section 2 contains details of the data sets used, including describing the creation of a WP reforecast data set. Section 3 describes the models in detail and the forecast verification procedure. In Sect. 4, we shall present the results and in Sect. 5, we draw some conclusions and make recommendations for future work.

2 Data

We use a Met Office WP classification called MO30 (Neal *et al.*, 2016). WPs in MO30 were defined by using simulated annealing to cluster 154 years (1850-2003) of daily MSLP anomaly fields into 30 distinct states. The data were extracted from the European and North Atlantic daily to multidecadal climate variability (EMULATE) data set (Ansell *et al.*, 2006) in the domain 30° W-20° E; 35°-70° N, with a spatial resolution of 5° latitude and longitude. These 30 WPs are therefore representative of the 30 most common patterns of daily atmospheric circulation over Europe and the North Atlantic (Fig. 1), and they were ordered such that WP1 is the most frequently occurring WP annually, while WP30 is the least frequent. A consequence of the clustering process and ordering is that the lower-numbered WPs have lower-magnitude MSLP anomalies and are more common in the summer than in the winter, and vice versa for the higher-numbered WPs (Richardson *et al.*, 2018a)(Neal *et al.*, 2016).

For this analysis, we have created a 20-year daily WP probabilistic reforecast data set. We use the sub-seasonal to seasonal (S2S) project (Vitart *et al.*, 2017) data archive, which, through ECMWF, hosts reforecast data for a multitude of variables and by a range of models from around the globe. In particular, we use ECMWF-EPS, which is a coupled atmosphere-ocean-sea-ice model with a lead-time of 46 days. The horizontal atmospheric resolution is roughly 16 km up to day 15 and 32 km beyond this. The model is run at 00Z, twice weekly (Mondays and Thursdays) and has 11 ensemble members for the reforecasts (compared to 51 members for the real-time forecasts). For further details, refer to the model webpage (ECMWF, 2017). We use daily reforecasts of MSLP between 02 January 1997 and 28 December 2016, inclusive, with the same domain and resolution as MO30. These fields are converted to forecast anomalies by removing a smoothed climatology and subsequently assigned to the closest matching MO30 WP via minimising the sum-of-squared differences. Both the MSLP climatology and the WP definitions are the same as those used by Neal *et al.* (2016) to ensure consistency. We compare this against an 'observed' WP time series to measure forecast skill. For this, WPs are assigned from 00Z SLP fields from the ERA-Interim reanalysis data set (Dee *et al.*, 2011) between 1979 and 2017. A consequence of assigning WPs using ERA-Interim compared to the EMULATE data set used in the original derivation of MO30 is that the historical frequencies of occurrence of the WPs differ. The same strongly seasonal behaviour is retained (lower-numbered WPs occurring more often in summer than higher-numbered WPs, and vice versa), but the annual frequencies are more evenly distributed across the WPs - there is no clear decrease in annual frequency as the WP number is increased (Figure 1).

As observed precipitation, we use the Met Office Hadley Centre UK Precipitation (HadUKP) data set (Alexander and Jones, 2000). For nine regions covering the UK, we use daily precipitation series from 1979 to 2017. We discretise the data into precipitation intervals ("bins") defined in Table 1; see Section 3.2 for further information. The large region sizes in HadUKP are suitable both for analyses of drought, which is typically considered a regional rather than localised event (Marsh *et al.*, 2007), and for MO30 because they correspond to the large-scale circulation patterns that the WPs represent. From the S2S archive, we extract ECMWF-EPS precipitation reforecasts for the same dates as the WP reforecast data set. The data have a resolution of 0.5° latitude and longitude; grid cells are assigned to whichever of the nine HadUKP regions the cell centres lie in (Fig. S1) and by taking the daily mean of all cells over each region, we produce a probabilistic reforecast data set of precipitation for each of the HadUKP regions. Then, we remove the three-monthly-mean bias of the forecasts compared to the observations for each region. The bias correction is done using leave-one-year-out cross validation. Finally, these data are discretised in the same way as the HadUKP data.

3 Methods

3.1 Weather pattern forecast models and verification procedure

For WP forecasts, we compare two models. The first is ECMWF-EPS, which we shall refer to as EPS-WP (in practice this is the WP reforecast data set discussed in the previous subsection). The second model is a 1000-member, first-order, nonhomogeneous Markov chain, with separate transition matrices for each month. This is similar to the Markov model used for a simulation study by Richardson *et al.* (2018b), who found it was able to reasonably replicate the observed frequencies of occurrences of the MO30 WPs. Full details of the Markov model are given in the supporting material.

To evaluate WP forecast skill we use the Jensen-Shannon divergence (JSD), suitable for measuring the distance between two probability distributions (Lin, 1991). It is based on information entropy, which is used to measure uncertainty. An information-theoretic approach to verification is not widespread, although there is some published research on the topic (Leung and North, 1990; Kleeman, 2002; Roulston and Smith, 2002; Ahrens and Walser, 2008; Weijs *et al.*, 2010; Weijs and Giesen, 2011). The JSD will be used to measure the forecast performance by quantifying the distance between distributions of the observed and forecast WP frequencies. The JSD is based on the Kullback-Leibler divergence (KLD) (Kullback and Leibler, 1951). Let P and Q be two discrete probability distributions. The KLD from Q to P is given by:

$$D_{KL}(P||Q) = -\sum_{i=1}^{I} P_i \log_2 \frac{Q_i}{P_i},$$

Equation 1

measured in bits (i.e. a binary unit of information). In our application $I = 30$, the number of WPs and $P = (p_{f,1}, \dots, p_{f,30})$ and $Q = (q_{f,1}, \dots, q_{f,30})$ are the vectors of observed and forecast WP relative frequencies, respectively. (Because these are relative frequencies, $\sum P = 1$ and $\sum Q = 1$.) As there would inevitably be some cases where the model predicts no occurrences of some WPs (i.e. when Q contains zeros), $D_{KL}(P||Q)$ will be undefined at times. Using the JSD avoids this problem; it is defined as:

$$D_{JSD}(P||Q) = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M),$$

Equation 2

where $M = (P + Q)/2$. Unlike the KLD, the JSD is symmetric i.e. $D_{JSD}(P||Q) \equiv D_{JSD}(Q||P)$. Also, $0 \leq D_{JSD}(P||Q) \leq 1$, with a score of zero indicating P and Q are the same (a perfect forecast). Equation 2 gives the JSD for a single forecast-event pair; to obtain the average JSD for all forecasts we take the mean of all forecast-event pairs. Skill is evaluated separately for each month, with the middle date of each forecast period used to assign the month. We calculate forecast skill for lead-times of 16, 31 and 46 days. We use the JSD to compare WP forecast skill of EPS-WP and the Markov model, considering each lead-time separately.

3.2 Precipitation and drought forecast models

We compare four models, three of which are forecast models, while one model is a perfect prognosis model. Fig. 2 shows a schematic of the procedure involved in generating forecasts from each model. All models are considered at the same lead-times as the WP predictions. Two of the forecast models are driven first by a WP component: EPS-WP and the Markov model described above. The perfect prognosis model, Perfect-WP, is used as an 'upper benchmark' with (future) observed WPs as input, rather than forecast WPs. It is an idealised model that cannot be used operationally, but it allows us to assess the potential

usefulness of WPs in precipitation and drought forecasting. Note that from here, any reference to drought refers specifically to meteorological drought.

Precipitation is estimated from the WP predictions (or observations in the case of Perfect-WP) by sampling from the conditional distributions of precipitation given each Era-Interim WP between 1979 and 2017. We process the daily HadUKP precipitation data by discretising into v bins with historical probabilities p_b for $b = 1, \dots, v$. Dry days form one bin and bin intervals increase for higher precipitation values (Table 1). This gives a discrete distribution of precipitation interval relative frequencies, $D(z)$, with conditional distributions for each WP given by $D(z|W = i)$, for $i = 1, \dots, 30$. We also define w summed precipitation intervals s_c for $c = 1, \dots, w$. Forecast probabilities of these summed intervals are derived from the WP forecast models as follows:

1. Set the ensemble member $e \in (e_1, \dots, e_N)$, where N_e is the number of ensemble members; time $t = 0$, the first day of the forecast, and then the predicted WP by ensemble member e at time t is $W_e(t) = i$ for $i = 1, \dots 30$.

2. Set $p_0 = 0$, calculate the probabilities $p_1, \dots p_m$ of each of the m daily precipitation bins from the discrete precipitation distribution that is conditional on $W_e(t)$ and on the 91-day windows centred on t (i.e. $t - 45, \dots, t + 45$) from every year except the current year. This last condition is equivalent to a leave-one-year-out cross-validation procedure.

3. Define the maximum value of each bin as $l_{p_b}, b = 1, \dots v$, with $l_{p_0} = 0$. Note that $l_{p_0} = l_{p_1} = 0$, ensuring zero precipitation days can be simulated.

4. Generate u random variables $p_k^* \sim U(0,1)$ for $k = 1, \dots, u$.

5. For each p_k^*, find the index q such that

$$\sum_{j=0}^{q} p_j < p_k^* < \sum_{j=0}^{q+1} p_j \, .$$

   Set $P_q = \sum_{j=0}^{q-1} p_j$ and $P_{q+1} = \sum_{j=0}^{q} p_j$, the cumulative probabilities of the bins adjacent to p_k^*.

6. Define the difference between the adjacent bins as $\alpha = P_{q+1} - P_q$ and the difference between the random number and the lower cumulative probability as $\beta = p_k^* - P_q$.

7. Estimate the precipitation value for each p_k^* as $r_k(t) = l_{p_q} + \frac{\beta}{\alpha}\left(l_{p_{q+1}} - l_{p_q}\right)$. We now have u predicted daily precipitation values at time t, $\boldsymbol{r}(t) = (r_1(t), \dots, r_n(t))$.

8. Set $t = t + 1$ and repeat steps 3 to 6 until the final day of the forecast, $t_{\max}$, is processed.

9. Sum the daily precipitation vectors and divide by the random-sample size $(\sum_\tau \boldsymbol{r}(t))/u$ for $\tau = 0, \dots, t_{\max}$.

10. Discretise according to the w summed precipitation bins $s_1, \dots s_w$ to obtain a distribution of relative frequencies for this ensemble member $\boldsymbol{f_e} = (f_1, \dots, f_w)$.

11. Set a new ensemble member $e^* \in \left(e_1, \dots e_{N_e}\right), e^* \neq e$ and repeat steps 2 to 10 until every ensemble member has been processed.

12. Sum each ensemble member's distribution of summed precipitation relative frequencies and divide by the number of ensemble members to obtain a final forecast probability distribution:

$$\boldsymbol{F} = \left(\sum_e \boldsymbol{f_e}\right) \Big/ N_e.$$

The number of ensemble members depends on the model. For EPS-WP, $N_e = 11$, i.e. the number of ensemble members of the ECMWF dynamical model. For the Markov model $N_e = 1000$. We set the number of samples drawn from each WP-precipitation conditional distribution as $u = 10{,}000$. The fourth model (the third forecast model) is the direct ECMWF-EPS precipitation forecasts (EPS-P), processed to provide probabilistic predictions of regional precipitation intervals as described earlier.

3.3 Precipitation forecast verification

To evaluate precipitation forecast performance we use the ranked probability score (RPS) (Epstein, 1969; Murphy, 1971). We express the RPS as the ranked probability skill score (RPSS) using

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS}_{\mathrm{ref}}},$$

Equation 3

Where $\mathrm{RPS}_{\mathrm{ref}}$ is the score of a climatological forecast, which in our case is the climatological event category (i.e. precipitation interval) relative frequencies (PC). A perfect score is achieved when $\mathrm{RPSS} = 1$, which is also the upper limit. Negative (positive) values indicate the forecast is performing worse (better) than $\mathrm{RPS}_{\mathrm{ref}}$.

3.4 Drought forecast verification

We evaluate model performance in predicting dichotomous drought/non-drought events. We define two classes of drought severity. The first class, mild drought, is when precipitation sums (over the length of the considered lead-time: either 16, 31 or 46 days) are below the 30.9$^{\mathrm{th}}$ percentile of the summed precipitation distribution. The second class is moderate drought, with such sums being below the 15.9$^{\mathrm{th}}$ percentile. These percentiles are calculated for each region and month using the whole data set from 1979 through 2017, and are chosen as they correspond to SPI values of -0.5 and -1, respectively.

3.4.1 The Brier Skill Score

We use three verification techniques to assess skill in predicting droughts. The first is the Brier Skill Score (BSS). The BSS is based on the Brier Score (BS) (Brier, 1950), which measures the mean-square error of probability forecasts for a dichotomous event, in this case the occurrence or non-occurrence of drought. The BS is converted to a relative measure, or skill score, by setting

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}},$$

Equation 4

where $\mathrm{BS}_{\mathrm{ref}}$ is the score of a reference forecast given by the quantiles associated with each drought threshold, 0.309 for mild drought and 0.159 for moderate drought. As with the RPSS, a perfect score is achieved when $\mathrm{BSS} = 1$ and negative (positive) values indicate the forecast is performing worse (better) than $\mathrm{BS}_{\mathrm{ref}}$.

3.4.2 Reliability diagrams – forecast reliability, resolution and sharpness

The BS can be decomposed into reliability, resolution and uncertainty terms (Murphy, 1973):

$$\mathrm{BS} = \mathrm{reliability} - \mathrm{resolution} + \mathrm{uncertainty},$$

Equation 5

enabling a more in-depth assessment of forecast model performance. Reliability diagrams offer a convenient way of visualising the first two of these terms (Wilks, 2011). These diagrams consist of two parts, which together show the full joint distribution of forecasts and observations. The first element is the calibration function, $g(o_1|p_i)$, for $i = 1, \dots n$, where o_1 indicates the event (here, a drought) occurring and the p_i are the forecast probabilities. The calibration function is visualised by plotting the event relative frequencies against the forecast probabilities and indicates how well calibrated the forecasts are. We split the forecast probabilities into 10 bins (subsamples) of 10% probability and the mean of all forecast probabilities in each bin is the value plotted on the diagrams (Bröcker and Smith, 2007). Points along the 1:1 line represent a well-calibrated, *reliable*, forecast, as event probabilities are equal to the forecast probabilities and suggest that we can interpret our forecasts at 'face value'. If the points are to the right (left) of the diagonal, the model is over-forecasting (under-forecasting) the number of drought events.

The forecast *resolution* can also be deduced from the calibration function. For a forecast with poor resolution, the event relative frequencies $g(o_1|p_i)$ only weakly depend on the forecast probabilities. This is reflected by a smaller difference between the calibration function and the horizontal line of the climatological event frequencies and suggests that the forecast is unable to resolve when a drought is more or less likely to occur than the climatological probability. Good resolution, on the other hand, means that the forecasts are able to distinguish different subsets of forecast occasions for which the subsequent event outcomes are different to each other.

The second element of reliability diagrams is the refinement distribution, $g(p_i)$. This expresses how confident the forecast models are by counting the number of times a forecast is issued in each probability bin. This feature is also called *sharpness*. A low-sharpness model would overwhelmingly predict drought at the climatological frequency, while a high-sharpness model would forecast drought at extreme high and low probabilities, reflecting its level of certainty with which a drought will or will not occur, independent of whether a drought actually does subsequently occur or not.

3.4.3 Relative operating characteristics

As a final diagnostic we use the relative operating characteristic (ROC) curve (Mason, 1982; Wilks, 2011), which visualises a model's ability to discriminate between events and non-events. Conditioned on the observations, the ROC curve may be considered a measure of potential usefulness – it essentially asks what the forecast is, given that a drought has occurred. The ROC curve plots the hit rate (when the model forecasts a drought and a drought subsequently occurs) against the false alarm rate (when the model forecasts a drought but a drought does not then occur). We compute the hit rate and false alarm rate for cumulative probabilities between 0% and 100% at intervals of 10%. A skilful forecast model will have a hit rate greater than a false alarm rate, and the ROC curve would therefore bow towards the top-left corner of the plot. The ROC curve of a forecast system with no skill would lie along the diagonal, as the hit rate and false alarm rate would be equal, meaning the forecast is no better than a random guess. The area under the ROC curve (AUC) is a useful scalar summary. AUC ranges between zero and one, with higher scores indicating greater skill.

4 Results

To reduce information overload, we do not show results for every combination of region, lead-time and drought class. Key results not shown will be conveyed via the text. We aggregate the precipitation results from monthly to three-month seasons for visual clarity and combine regional results for the ROC and reliability diagrams for the same reason.

4.1 WP forecasts

We find that EPS-WP is more skilful at predicting the WPs than the Markov model for every month and every lead-time, although the difference in skill between the two models decreases as the lead-time increases. The skill difference between models is much larger for a lead-time of 16 days compared to a lead-time of 46 days (Fig. 3). For a 46-day lead-time, the

difference in skill is negligible for May through October; in fact, these months have the smallest differences in JSD for all lead-times. This is presumably because the summer months are associated with fewer WPs compared to winter (Richardson *et al.*, 2018a), resulting in a more skilful Markov model due to higher transition probabilities.

An interesting result is how JSD scores for Markov decrease as the lead-time increases (Fig. 3), suggesting an improvement in skill with lead-time. This is the opposite of the expected (and usual) effect. The Markov model predicts WPs using the one-day transition probabilities, and its ensemble members therefore diverge very quickly, resulting in a distribution of predicted WPs that looks similar to the climatological WP distribution for all lead-times. For a 16-day forecast, the observed WP distribution of the corresponding 16 days will generally be less similar to the climatological WP distribution than for 31-day forecasts, and less similar still than for 46-day forecasts. For instance, at a 16-day lead, only 16 unique WPs could form the observed distribution, whereas Markov is capable of predicting all possible WPs across its 1000 members at this lead. As the JSD measures the distance between these probability distributions, it tends to score the differences between these distributions as more similar (a smaller divergence) for longer lead-times. This means the JSD is perhaps not appropriate as a verification metric in an operational sense, but is noteworthy for highlighting the behaviour of the Markov model.

We could have assessed model skill in predicting the WPs using more common metrics such as the BS, which could measure the hit/miss ratio for each WP at each lead-time. However, the focus of this paper is on multi-week precipitation (and drought) totals, so we are not particularly interested in the models' ability to predict the timing of a WP, only whether they are able to capture the distribution of the WP frequencies of occurrence. It is likely that using the BS would show that EPS-WP and Markov skill decreases with lead-time, as was the case for a WP classification derived from MO30 by Neal *et al.* (2016).

4.2 Precipitation forecasts

We first discuss the skill of the three true forecast models, EPS-WP, EPS-P and Markov. For the most part, all three models are more skilful than climatology independent of season and lead-time, with greater skill in autumn and winter compared to spring and summer (Figs. 4 and 5). For a 16-day lead-time, there is little to choose between EPS-WP and EPS-P, except in ES, for which the latter model is less skilful than climatology in winter and spring (Fig. 4). Markov is the least skilful model at this lead, offering only a marginal improvement on climatology (Fig. 4). The skill of EPS-WP and EPS-P reduces when a 31-day lead is considered, bringing their skill more in line with Markov (Fig. S2). At a 46-day lead the differences are starker, with EPS-P notably less skilful than EPS-WP, Markov and climatology for many regions in summer and, especially, spring (Fig. 5). These results are, however, still only marginally superior to climatology. EPS-WP has greater skill than EPS-P at this lead-time in winter and autumn for NS, NI, CEE and SWE, although the magnitudes of these differences are small (Fig. 5). There is little evidence of coherent regional variability in model skill, except perhaps a tendency for EPS-P to score more highly for western regions in spring and summer at a 16-day lead-time (Fig. 4). Despite low skill relative to climatology at longer lead-times, there is clearly some benefit to using the WP-based models (particularly EPS-WP) for certain regions and seasons.

The potential usefulness of such approaches is highlighted by the performance of Perfect-WP. Unsurprisingly, this model is almost uniformly the most skilful model for all regions, seasons and lead-times (Figs. 4, 5 and S2). The gains in skill for this model over the other three models are most pronounced during winter and autumn and especially for longer lead-times. Skill is greatest for most western regions (NS, NI, NWE and SWE) and lowest for eastern regions ES, NEE and SEE, together with SS (Fig. 5). Perfect-WP is obviously not practical, but the results serve to show that WPs are a potentially useful tool in medium-range precipitation forecasting.

4.3 Meteorological drought forecasts

4.3.1 Forecast accuracy

Forecast accuracy is typically lower for mild drought (total precipitation over 16, 31, or 46 days below the 30.9th percentile) than for precipitation, and lower still for moderate drought (total precipitation below the 15.9th percentile). The regional and lead-time differences in precipitation skill are also evident for drought, with higher skill at shorter leads and during winter and autumn (Figs. 6, 7 and S3). Results for mild drought are not shown as they generally lie in-between those for precipitation (Figs. 4, 5 and S2) and moderate drought (Figs. 6, 7 and S3). Markov again has the poorest skill, with a climatology forecast preferable for many combinations of region and lead-time. EPS-P is either equal or more skilful than EPS-WP at a 16-day lead (Fig. 6), and during spring for longer leads (Figs. 7 and S3). Conversely, EPS-WP outperforms EPS-P during summer at the longer two lead-times, although a climatology forecast would be just as, if not more skilful. As with precipitation forecasts, any gain in skill using EPS-WP over EPS-P in winter and autumn at longer leads is marginal, with both models showing more skill than climatology (Figs 7 and S3).

Skill, where present, is undeniably modest, but the relatively high skill of Perfect-WP in some regions and seasons again shows the potential predictability of drought using WP methods. Compared to precipitation forecasts, skill for Perfect-WP is notably lower for spring and summer, with climatology often a competitive forecast method at a 46-day lead-time (Fig. 7). For winter and autumn, however, the skill is reasonable UK-wide, and particularly high during winter in NS and NI (Fig. 7). The same east-west skill split is present for moderate drought as it was for precipitation, with some western regions benefitting from higher skill than eastern region (Fig. 7).

4.3.2 Relative operating characteristics

All models are better able to discriminate between drought and non-drought events than random chance, with Perfect-WP the most able and Markov the least able, subject to similar caveats regarding lead-time and season as for the BSS and RPSS results. During summer and spring, EPS-P has the highest AUC of any of the three forecast models (Figs. 8 and 9), and for a 16-day lead-time scores similarly to Perfect-WP (not shown). On the other hand, EPS-WP is the best discriminator during winter and autumn at a 46-day lead-time, although the magnitude of the differences is small (Figs. 8 and 9). Markov is consistently the least suitable model for predicting drought according to the ROC curve, although still represents a better method of doing so than random chance.

A use of the ROC curve is to provide end-users with information on how to apply the considered forecast models. As the plotted points on each curve indicate the hit rate and false alarm rate associated with predicting droughts at each probability interval, they can be used to make an informed decision in selecting a probability threshold for issuing a drought forecast. For example, should a forecaster choose to issue a moderate drought warning in winter at a 10% probability level and 46-day lead-time (Fig. 9), then they would expect EPS-WP to achieve a hit rate over double that of the false alarm rate (~55% and ~20%, respectively). EPS-P, meanwhile, shows a slightly lower hit rate and similar false alarm rate (~50% and ~20%). The idealised benchmark model (Perfect-WP) achieves an outstanding score – an over 70% hit rate compared to a <10% false alarm rate. For mild drought, a 20% probability threshold for EPS-WP and EPS-P achieves at least 60% hit rates at all lead-times, whereas for moderate drought, this threshold will only achieve such hit rates at a 16-day lead-time during winter and autumn (EPS-P also achieves this rate for spring and summer; not shown) and during autumn for all lead-times. In general, it appears that these low probability thresholds yield the best compromise between hits and false alarms, although in practice, the costs (e.g. financial) associated with false alarms and missed events will determine how responders use these probabilities.

4.3.3 Forecast reliability, resolution and sharpness

EPS-WP is the most reliable forecast model (i.e. excluding Perfect-WP), and while all three WP-driven forecast models tend to under-forecast droughts, EPS-P only does so for lower probability thresholds, with the higher thresholds resulting in this model over-forecasting. This is particularly true for shorter lead-times and during winter, although is still clear for 31-day

lead-times in some seasons (Figs. 10 and 11). Sometimes EPS-WP follows the same pattern as EPS-P and over-forecasts drought occurrence for higher predicted probabilities (e.g. Figs. 10c, e, g and 11c). However, the total number of forecasts issued in these intervals is generally smaller than for EPS-P, as the refinement distributions show most clearly for mild drought (Fig. 10). This means the corresponding points of the calibration function are less reliable for EPS-WP (and Markov) due to smaller sample sizes (Bröcker and Smith, 2007). In fact, all three WP-based models have occasions when there are no issued forecasts with certain probabilities. These are high probabilities for Perfect-WP and EPS-WP (Figs. 11c and e) but can be as low as between 30% and 40% for Markov (Figs. 11e and g). As such, although EPS-WP appears the most reliable model from looking only at the calibration function, there is less certainty of this fact for moderate drought and for higher forecast probabilities. This erratic behaviour of the conditional event relative frequencies is most obvious in Fig. 11c and is explained by the very low sample sizes of forecasts issued with anything but a small probability (Fig. 11e) (Wilks, 1995). An interesting result is that forecasts from EPS-WP are more reliable than from Perfect-WP when the predicted drought probabilities are below 80% for mild drought (Fig. 10) and 60% for moderate drought (except in spring; Fig. 11), despite having lower accuracy (e.g. Fig. 6). As a more skilful BSS is composed of smaller reliability and larger resolution terms (Kharin and Zwiers, 2003), it follows that the resolution of Perfect-WP is sufficiently large to overcome the larger reliability term compared to EPS-WP and yield an overall more accurate forecast model. However, for drought forecasts issued with higher probabilities, EPS-WP is the less reliable model, under- or over-forecasting drought (depending on the season) more than Perfect-WP. These under- or over-forecasting biases must be taken into account by an operational forecaster using these models.

A key difference apparent from the calibration function relates to the ability of the models to identify subsets of forecast situations where the subsequent event relative frequencies are different, i.e. the forecast resolution. A fairly consistent feature across all lead-times and drought classes is the poorer resolution of EPS-P, particularly obvious in summer (Figs. 10e and 11e), with the conditional event relative frequencies quite clearly closer to the climatological average compared to the other models. This should be considered in conjunction with the sharpness of the forecast, which is relatively high for this model as shown by the numbers of issued extreme probabilities, particularly those in the upper-tail (Figs. 10f and 11f). This combination of poor resolution and high sharpness indicates "overconfidence" (Wilks, 2011) – on the occasions that EPS-P issues a forecast indicating the likelihood of a drought is very high, the actual likelihood of a drought subsequently occurring is lower. To compensate for this overconfidence, a user would adjust the probabilities to be less extreme to make the forecasts more reliable.

We can compare these refinement distributions to those of the Markov model, which exhibits low sharpness, overwhelmingly predicting droughts at the climatological frequency (second column of Figs. 10 and 11). This means that the Markov model is not a useful operational tool in these situations, as similar forecasts could be obtained simply by using the climatological drought frequency. The refinement distributions for EPS-WP show that for mild drought in winter and spring and for moderate drought in all seasons, the model predicts droughts with low probabilities the majority of the time (Figs. 10b, d and 11b, d, f, h). For mild drought in summer and autumn, however, this model mostly issues forecasts close to the climatological frequency, although not nearly as regularly as the Markov model (Fig. 10f, h). As with adjusting for bias, a forecaster can use model resolution and sharpness when assessing drought forecast probabilities output by a model.

5 Discussion and conclusions

We have compared the performance of a dynamical forecast system (EPS-WP) and a first-order Markov model in predicting WP occurrences over a range of lead-times, showing that the dynamical model is always more skilful, although the difference in skill reduces with lead-time. From these WP predictions, we derived precipitation and meteorological drought forecasts and compared them to direct precipitation and drought predictions from the dynamical system (EPS-P). We compared two levels of drought: mild drought, when the total precipitation over the lead-time (16, 31 or 46 days) was below the 30.9^{th} percentile climatology, and moderate drought, when the total precipitation over the lead-time was below the 15.9^{th} percentile. Overall, forecast models were found to be more skilful during winter and autumn, particular for longer lead-times. The Markov model

tended to be the least skilful, especially when predicting drought. Differences in skill between EPS-P and EPS-WP were typically small, with RPSS, BSS and ROC results not highlighting a clear winner. However, we demonstrated the potential in improving WP forecasts further by showing that an idealised, perfect prognosis model (Perfect-WP) would provide much more skilful precipitation and drought forecasts, with high hit rates and low false alarm rates.

From assessing reliability diagrams, we found that WP-based models only issue binary drought forecasts with either very low probabilities or probabilities close to the climatological average. In particular, there is little to gain in using the Markov model in mild drought prediction over the climatological frequency, as it tends to issue drought forecasts with this probability anyway. EPS-P has the highest sharpness, predicting drought occurrence with a wide range of probabilities. In particular, it issues greater numbers of high-probability drought forecasts compared to WP-based methods. However, this model also has poor resolution, indicating it is an overconfident forecast model. Overall, drought forecasts issued by EPS-WP are the most reliable, i.e. the forecast probabilities are most similar to the subsequent event probabilities (they "mean what they say") (Wilks, 2011). Perfect-WP tends to under-forecast the number of drought events, while EPS-P over-forecasts drought events, particularly for moderate drought. These reliability diagrams are therefore useful to aid users in adjusting for an over- or under-forecasting bias.

The higher skill of EPS-WP during winter (and possibly autumn) is probably due to the typically higher skill that medium- to long-range dynamical forecast systems have in predicting atmospheric variables in this season compared to other seasons (Scaife *et al.*, 2014; MacLachlan *et al.*, 2015; Neal *et al.*, 2016; Arnal *et al.*, 2018). In fact, by forecasting a set of eight WPs derived from MO30, Neal *et al.* (2016) found that ECMWF-EPS exhibited greater skill in winter than summer. Furthermore, the relationship between the NAO (which is the primary mode of North Atlantic/European atmospheric circulation) and precipitation is stronger in this season (Hurrell and Deser, 2009; Lavers *et al.*, 2010; Svensson *et al.*, 2015). This is particularly true for western regions (Jones *et al.*, 2013; Svensson *et al.*, 2015; van Oldenborgh *et al.*, 2015; Hall and Hanna, 2018), which potentially explains the greater skill of precipitation and drought forecasting using observed WPs (Perfect-WP). The regional variations in skill of this model imply that MO30 is not as suited for representing precipitation in the east. Perhaps this is because the WPs are more closely related to the NAO in this season compared to other teleconnection patterns. As Hall and Hanna (2018) showed, the NAO is not the only important teleconnection pattern influencing UK precipitation.

By analysing the skill of an idealised 'forecast' model that assumes perfect WP predictions, we have demonstrated the potential for using WP forecasts to derive precipitation and drought predictions. The skill of this model during winter and autumn suggest that the processes between the WPs and precipitation are well represented in these seasons. The lesser skill of EPS-WP and Markov, then, is a result of poor prediction of the WPs. A focus on improving the skill of the WP forecasts could be the most useful route to improving precipitation and drought predicting skill. Currently, dynamical models such as the ECMWF system used here represent the best method of predicting WPs. Moreover, the ECMWF reforecast data used here had 11 ensemble members, whereas the operational forecasts are run with 51 members. Therefore, an operationalised version of the models might improve forecast skill or better represent uncertainty, although this is also true for precipitation forecasts direct from the model. A useful piece of further research would be to assess the forecast skill of other models, and multi-model ensembles, at predicting MO30 WPs or other WP classification systems. Another potential method to improve precipitation and drought forecast skill would be to alter the process by which precipitation is estimated from the WPs. Here we have sampled from the entire conditional distribution of precipitation given the WP and season, but this may not be the optimal way of estimation. It is possible that other factors influence the precipitation from WPs, such as slowly-varying atmospheric and oceanic processes. For example, it would be interesting to see if conditioning the distributions further on the state of the NAO index, or some North Atlantic SST index, and sampling precipitation from these, would improve forecast skill. This is potentially most useful in predicting moderate drought, for which skill from current models is lower than for mild drought.

Code availability

The code is only available locally with DR. Please contact the corresponding author for any queries regarding sharing the code.

Data availability

Met Office EMULATE MSLP data can be found at https://www.metoffice.gov.uk/hadobs/emslp/; ERA-Interim data at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim; ECMWF EPS hindcast data at https://apps.ecmwf.int/datasets/data/s2s/ and Met Office HadUKP data at https://www.metoffice.gov.uk/hadobs/hadukp/.

Author contribution

D. Richardson was the primary designer of the experiment, developed the model code, produced the figures and wrote the manuscript. H. Fowler, C. Kilsby, R. Neal and R. Dankers contributed to the design of the experiment and provided input into figure and text editing.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We thank two reviewers for their insightful comments and suggestions that improved the quality of this article. This work was part of a NERC funded Postgraduate Research Student Studentship NE/L010518/1. H.J.F. is funded by the Wolfson Foundation and the Royal Society as a Royal Society Wolfson Research Merit Award (WM140025) holder. H.J.F. acknowledges support from the INTENSE project supported by the European Research Council (grant ERC-2013-CoG-617329).

List of references

Ahrens, B. and Walser, A. (2008) 'Information-Based Skill Scores for Probabilistic Forecasts', *Monthly Weather Review*, 136(1), pp. 352-363.

Alexander, L.V. and Jones, P.D. (2000) 'Updated Precipitation Series for the U.K. and Discussion of Recent Extremes', *Atmospheric Science Letters*, 1(2), pp. 142-150.

Ansell, T.J., Jones, P.D., Allan, R.J., Lister, D., Parker, D.E., Brunet, M., Moberg, A., Jacobeit, J., Brohan, P., Rayner, N.A., Aguilar, E., Alexandersson, H., Barriendos, M., Brandsma, T., Cox, N.J., Della-Marta, P.M., Drebs, A., Founda, D., Gerstengarbe, F., Hickey, K., Jónsson, T., Luterbacher, J., Ø, N., Oesterle, H., Petrakis, M., Philipp, A., Rodwell, M.J., Saladie, O., Sigro, J., Slonosky, V., Srnec, L., Swail, V., García-Suárez, A.M., Tuomenvirta, H., Wang, X., Wanner, H., Werner, P., Wheeler, D. and Xoplaki, E. (2006) 'Daily Mean Sea Level Pressure Reconstructions for the European–North Atlantic Region for the Period 1850–2003', *Journal of Climate*, 19(12), pp. 2717-2742.

Arnal, L., Cloke, H.L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B. and Pappenberger, F. (2018) 'Skilful seasonal forecasts of streamflow over Europe?', *Hydrol. Earth Syst. Sci.*, 22(4), pp. 2057-2072.

Baker, L.H., Shaffrey, L.C. and Scaife, A.A. (2018) 'Improved seasonal prediction of UK regional precipitation using atmospheric circulation', *International Journal of Climatology*, 38, pp. 437-453.

Bárdossy, A. and Filiz, F. (2005) 'Identification of flood producing atmospheric circulation patterns', *Journal of Hydrology*, 313(1–2), pp. 48-57.

Brier, G.W. (1950) 'Verification of forecasts expressed in terms of probability', *Monthly Weather Review*, 78(1), pp. 1-3.

Brigode, P., Gérardin, M., Bernardara, P., Gailhard, J. and Ribstein, P. (2018) 'Changes in French weather pattern seasonal frequencies projected by a CMIP5 ensemble', *International Journal of Climatology*, 38(10), pp. 3991-4006.

Bröcker, J. and Smith, L., A. (2007) 'Increasing the Reliability of Reliability Diagrams', *Weather and Forecasting*, 22(3), pp. 651-661.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G. and Vitart, F. (2007) 'The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System)', *Quarterly Journal of the Royal Meteorological Society*, 133(624), pp. 681-695.

Cuo, L., Pagano, T.C. and Wang, Q.J. (2011) 'A Review of Quantitative Precipitation Forecasts and Their Use in Short- to Medium-Range Streamflow Forecasting', *Journal of Hydrometeorology*, 12(5), pp. 713-728.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N. and Vitart, F. (2011) 'The ERA-Interim reanalysis: configuration and performance of the data assimilation system', *Quarterly Journal of the Royal Meteorological Society*, 137(656), pp. 553-597.

Dutra, E., Di Giuseppe, F., Wetterhall, F. and Pappenberger, F. (2013) 'Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index', *Hydrol. Earth Syst. Sci.*, 17(6), pp. 2359-2373.

ECMWF (2017) *ECMWF Model Description CY43R1 [Online]*. Available at: https://confluence.ecmwf.int/display/S2S/ECMWF+Model+Description+CY43R1 (Accessed: 03/06/2018).

Epstein, E., S. (1969) 'A Scoring System for Probability Forecasts of Ranked Categories', *Journal of Applied Meteorology*, 8(6), pp. 985-987.

Ferranti, L., Corti, S. and Janousek, M. (2015) 'Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector', *Quarterly Journal of the Royal Meteorological Society*, 141(688), pp. 916-924.

Golding, B.W. (2000) 'Quantitative precipitation forecasting in the UK', *Journal of Hydrology*, 239(1), pp. 286-305.

Hall, R.J. and Hanna, E. (2018) 'North Atlantic circulation indices: links with summer and winter UK temperature and precipitation and implications for seasonal forecasting', *International Journal of Climatology*, 38(S1), pp. e660-e677.

Hannaford, J., Lloyd-Hughes, B., Keef, C., Parry, S. and Prudhomme, C. (2011) 'Examining the large-scale spatial coherence of European drought using regional indicators of precipitation and streamflow deficit', *Hydrological Processes*, 25(7), pp. 1146-1162.

Hay, L.E., McCabe, G.J., Wolock, D.M. and Ayers, M.A. (1991) 'Simulation of precipitation by weather type analysis', *Water Resources Research*, 27(4), pp. 493-501.

Hay, L.E., McCabe, G.J., Wolock, D.M. and Ayers, M.A. (1992) 'Use of weather types to disaggregate general circulation model predictions', *Journal of Geophysical Research: Atmospheres*, 97(D3), pp. 2781-2790.

Hurrell, J.W. and Deser, C. (2009) 'North Atlantic climate variability: The role of the North Atlantic Oscillation', *Journal of Marine Systems*, 78(1), pp. 28-41.

Huth, R., Beck, C., Philipp, A., Demuzere, M., Ustrnul, Z., Cahynová, M., Kyselý, J. and Tveito, O.E. (2008) 'Classifications of Atmospheric Circulation Patterns', *Annals of the New York Academy of Sciences*, 1146(1), pp. 105-152.

Jones, M.R., Fowler, H.J., Kilsby, C.G. and Blenkinsop, S. (2013) 'An assessment of changes in seasonal and annual extreme rainfall in the UK between 1961 and 2009', *International Journal of Climatology*, 33(5), pp. 1178-1194.

Kendon, M., Marsh, T. and Parry, S. (2013) 'The 2010–2012 drought in England and Wales', *Weather*, 68(4), pp. 88-95.

Kharin, V.V. and Zwiers, F.W. (2003) 'Improved Seasonal Probability Forecasts', *Journal of Climate*, 16(11), pp. 1684-1701.

Kleeman, R. (2002) 'Measuring Dynamical Prediction Utility Using Relative Entropy', *Journal of the Atmospheric Sciences*, 59(13), pp. 2057-2072.

Kullback, S. and Leibler, R.A. (1951) 'On Information and Sufficiency', *Ann. Math. Statist.*, 22(1), pp. 79-86.

Lavaysse, C., Vogt, J. and Pappenberger, F. (2015) 'Early warning of drought in Europe using the monthly ensemble system from ECMWF', *Hydrol. Earth Syst. Sci.*, 19(7), pp. 3273-3286.

Lavaysse, C., Vogt, J., Toreti, A., Carrera, M.L. and Pappenberger, F. (2018) 'On the use of weather regimes to forecast meteorological drought over Europe', *Nat. Hazards Earth Syst. Sci.*, 18(12), pp. 3297-3309.

Lavers, D., Prudhomme, C. and Hannah, D.M. (2010) 'Large-scale climate, precipitation and British river flows: Identifying hydroclimatological connections and dynamics', *Journal of Hydrology*, 395(3), pp. 242-255.

Lavers, D.A., Pappenberger, F. and Zsoter, E. (2014) 'Extending medium-range predictability of extreme hydrological events in Europe', *Nature Communications*, 5, p. 5382.

Lavers, D.A., Waliser, D.E., Ralph, F.M. and Dettinger, M.D. (2016) 'Predictability of horizontal water vapor transport relative to precipitation: Enhancing situational awareness for forecasting western U.S. extreme precipitation and flooding', *Geophysical Research Letters*, 43(5), pp. 2275-2282.

Leung, L.-Y. and North, G., R. (1990) 'Information Theory and Climate Prediction', *Journal of Climate*, 3(1), pp. 5-14.

Lin, J. (1991) 'Divergence measures based on the Shannon entropy', *IEEE Transactions on Information Theory*, 37(1), pp. 145-151.

MacLachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A., Gordon, M., Vellinga, M., Williams, A., Comer, R.E., Camp, J., Xavier, P. and Madec, G. (2015) 'Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system', *Quarterly Journal of the Royal Meteorological Society*, 141(689), pp. 1072-1084.

Marsh, T., Cole, G. and Wilby, R. (2007) 'Major droughts in England and Wales, 1800–2006', *Weather*, 62(4), pp. 87-93.

Marsh, T.J. (1995) 'The 1995 drought - a water resources review in the context of the recent hydrological instability', *LTA*, 155(47), p. 149.

Mason, I. (1982) 'A model for assessment of weather forecasts', *Australian Meteorological Magazine*, 30(4), pp. 291-303.

McKee, T.B., Doesken, N.J. and Kleist, J. (1993) 'The relationship of drought frequency and duration to time scales', *Proceedings of the 8th Conference on Applied Climatology*. American Meteorological Society Boston, MA. Available at: http://clima1.cptec.inpe.br/~rclima1/pdf/paper_spi.pdf.

Murphy, A., H. (1973) 'A New Vector Partition of the Probability Score', *Journal of Applied Meteorology*, 12(4), pp. 595-600.

Murphy, A., H. (1971) 'A Note on the Ranked Probability Score', *Journal of Applied Meteorology*, 10(1), pp. 155-156.

Mwangi, E., Wetterhall, F., Dutra, E., Di Giuseppe, F. and Pappenberger, F. (2014) 'Forecasting droughts in East Africa', *Hydrol. Earth Syst. Sci.*, 18(2), pp. 611-620.

Neal, R., Dankers, R., Saulter, A., Lane, A., Millard, J., Robbins, G. and Price, D. (2018) 'Use of probabilistic medium- to long-range weather-pattern forecasts for identifying periods with an increased likelihood of coastal flooding around the UK', *Meteorological Applications*, 25(4), pp. 534-547.

Neal, R., Fereday, D., Crocker, R. and Comer, R.E. (2016) 'A flexible approach to defining weather patterns and their application in weather forecasting over Europe', *Meteorological Applications*, 23(3), pp. 389-400.

Nigro, M., A. , Cassano, J., J. and Seefeldt, M., W. (2011) 'A Weather-Pattern-Based Approach to Evaluate the Antarctic Mesoscale Prediction System (AMPS) Forecasts: Comparison to Automatic Weather Station Observations', *Weather and Forecasting*, 26(2), pp. 184-198.

Richardson, D., Fowler, H.J., Kilsby, C.G. and Neal, R. (2018a) 'A new precipitation and drought climatology based on weather patterns', *International Journal of Climatology*, 38(2), pp. 630-648.

Richardson, D., Kilsby, C.G., Fowler, H.J. and Bárdossy, A. (2018b) 'Weekly to multi-month persistence in sets of daily weather patterns over Europe and the North Atlantic Ocean', *International Journal of Climatology*.

Richardson, D., Neal, R. and Dankers, R. (in review) 'Early warning of potential extreme precipitation events: a weather pattern approach', *Submitted for review to Meteorological Applications*.

Rodda, J.C. and Marsh, T.J. (2011) *The 1975-76 Drought - a contemporary and retrospective review*. [Online]. Available at: http://www.ceh.ac.uk/data/nrfa/nhmp/other_reports/CEH_1975-76_Drought_Report_Rodda_and_Marsh.pdf.

Roulston, M., S. and Smith, L., A. (2002) 'Evaluating Probabilistic Forecasts Using Information Theory', *Monthly Weather Review*, 130(6), pp. 1653-1660.

Saha, S., Shrinivas, M., Xingren, W., Jiande, W., Sudhir, N., Patrick, T., David, B., Yu-Tai, H., Hui-ya, C., Mark, I., Michael, E., Jesse, M., Rongqian, Y., Malaquías Peña, M., Huug van den, D., Qin, Z., Wanqiu, W., Mingyue, C. and Emily, B. (2014) 'The NCEP Climate Forecast System Version 2', *Journal of Climate*, 27(6), pp. 2185-2208.

Scaife, A.A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R.T., Dunstone, N., Eade, R., Fereday, D., Folland, C.K., Gordon, M., Hermanson, L., Knight, J.R., Lea, D.J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A.K., Smith, D., Vellinga, M., Wallace, E., Waters, J. and Williams, A. (2014) 'Skillful long-range prediction of European and North American winters', *Geophysical Research Letters*, 41(7), pp. 2514-2519.

Smith, D.M., Scaife, A.A. and Kirtman, B.P. (2012) 'What is the current state of scientific knowledge with regard to seasonal and decadal forecasting?', *Environmental Research Letters*, 7(1).

Svensson, C., Brookshaw, A., Scaife, A.A., Bell, V.A., Mackay, J.D., Jackson, C.R., Hannaford, J., Davies, H.N., Arribas, A. and Stanley, S. (2015) 'Long-range forecasts of UK winter hydrology', *Environmental Research Letters*, 10(6), p. 064006.

van Oldenborgh, G.J., Stephenson, D.B., Sterl, A., Vautard, R., Yiou, P., Drijfhout, S.S., von Storch, H. and van den Dool, H. (2015) 'Drivers of the 2013/14 winter floods in the UK', *Nature Climate Change*, 5, p. 490.

Vitart, F. (2014) 'Evolution of ECMWF sub-seasonal forecast skill scores', *Quarterly Journal of the Royal Meteorological Society*, 140(683), pp. 1889-1899.

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(1), pp. 163-173.

Vitart, F., Buizza, R., Alonso Balmaseda, M., Balsamo, G., Bidlot, J.-R., Bonet, A., Fuentes, M., Hofstadler, A., Molteni, F. and Palmer, T.N. (2008) 'The new VarEPS-monthly forecasting system: A first step towards seamless prediction', *Quarterly Journal of the Royal Meteorological Society*, 134(636), pp. 1789-1799.

Vuillaume, J.-F. and Herath, S. (2017) 'Improving global rainfall forecasting with a weather type approach in Japan', *Hydrological Sciences Journal*, 62(2), pp. 167-181.

Walker, G.T. and Bliss, E.W. (1932) 'World Weather V', *Memoirs of the Royal Meteorological Society*, 4(36), pp. 53-84.

Wedgbrow, C.S., Wilby, R. and Fox, H.R. (2005) 'Experimental seasonal forecasts of low summer flows in the River Thames, UK, using Expert Systems', *Climate Research*, 28(2), pp. 133-141.

Wedgbrow, C.S., Wilby, R.L., Fox, H.R. and O'Hare, G. (2002) 'Prospects for seasonal forecasting of summer drought and low river flow anomalies in England and Wales', *International Journal of Climatology*, 22(2), pp. 219-236.

Weijs, S., V. and Giesen, N.v.d. (2011) 'Accounting for Observational Uncertainty in Forecast Verification: An Information-Theoretical View on Forecasts, Observations, and Truth', *Monthly Weather Review*, 139(7), pp. 2156-2162.

Weijs, S., V., Nooijen, R.v. and Giesen, N.v.d. (2010) 'Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–Uncertainty Decomposition', *Monthly Weather Review*, 138(9), pp. 3387-3399.

Wilby, R.L. (1994) 'Stochastic weather type simulation for regional climate change impact assessment', *Water Resources Research*, 30(12), pp. 3395-3403.

Wilby, R.L. (1998) 'Modelling low-frequency rainfall events using airflow indices, weather patterns and frontal frequencies', *Journal of Hydrology*, 212–213, pp. 380-392.

Wilks, D.S. (1995) 'Chapter 7 Forecast verification', in Wilks, D.S. (ed.) *International Geophysics*. Academic Press, pp. 233-283.

Wilks, D.S. (2011) 'Chapter 8 - Forecast Verification', in Wilks, D.S. (ed.) *International Geophysics*. Academic Press, pp. 301-394.

Yoon, J.-H., Mo, K. and Wood, E.F. (2012) 'Dynamic-Model-Based Seasonal Prediction of Meteorological Drought over the Contiguous United States', *Journal of Hydrometeorology*, 13(2), pp. 463-482.

Yuan, X. and Wood, E.F. (2013) 'Multimodel seasonal forecasting of global drought onset', *Geophysical Research Letters*, 40(18), pp. 4900-4905.

Daily precipitation		**Total 16-, 31- and 46-day precipitation**	
p_b	**Range of precipitation, x, (mm)**	s_c	**Range of summed precipitation, y, (mm)**
p_1	0	s_1	$0 < y \leq 10$
p_2	$0 < x \leq 1$	s_2	$10 < y \leq 20$
…	Intervals of 1 mm	…	Intervals of 10 mm
p_{11}	$9 < x \leq 10$	s_{25}	$240 < y \leq 250$
p_{12}	$10 < x \leq 15$	s_{26}	$250 < y \leq 300$
p_{13}	$15 < x \leq 20$	…	Intervals of 50 mm
p_{14}	$20 < x \leq 30$	s_{30}	$300 < y \leq 450$
…	Intervals of 10 mm		
p_v	$90 < x \leq 100$		

Table 1: Range of daily precipitation, x, for each bin p_b and of 16-, 31- and 46-day total precipitation, y, for each bin s_c.

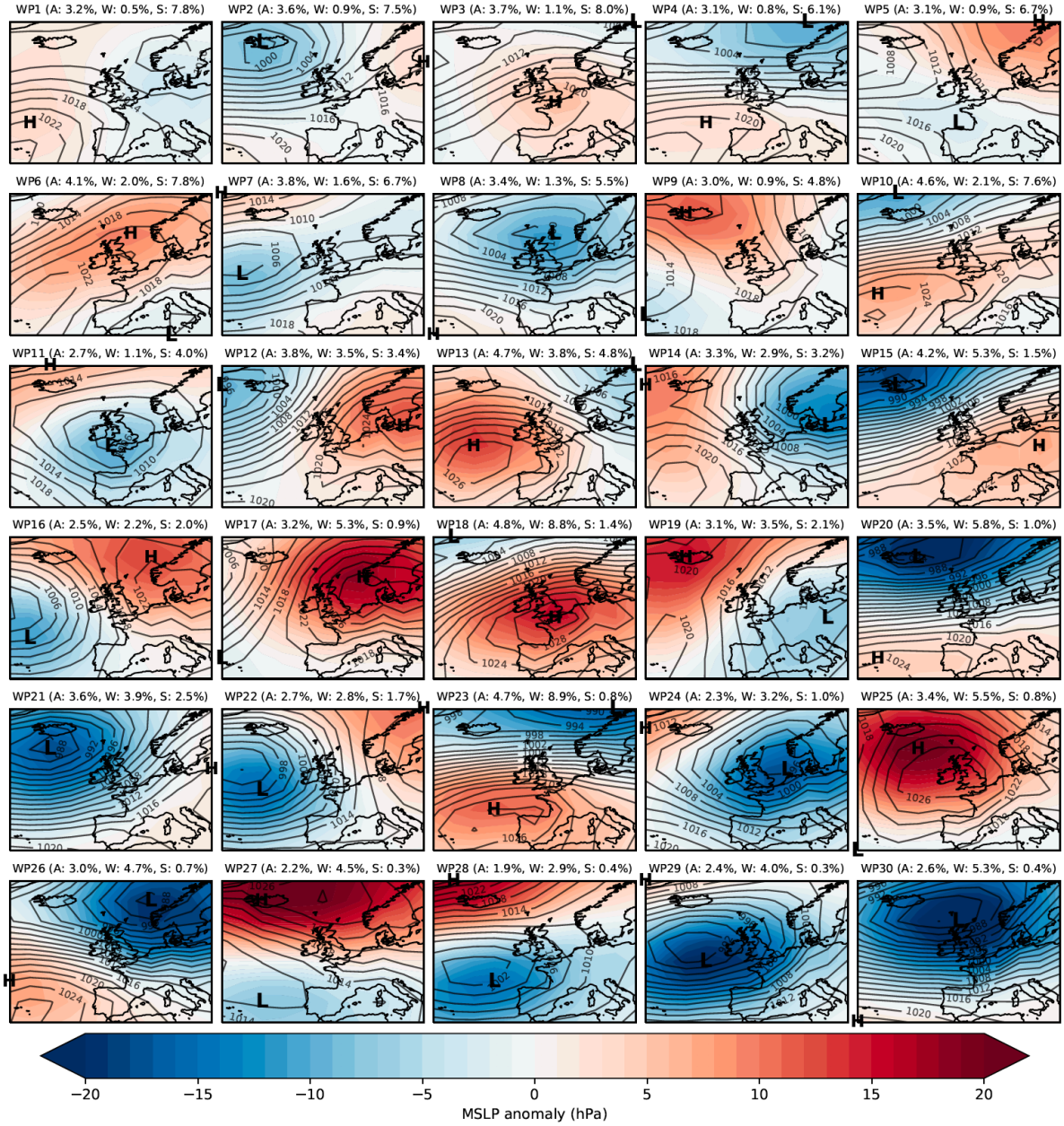

Figure 1: Weather pattern (WP) definitions according to mean sea-level pressure (MSLP) anomalies (hPa). The black contours are isobars showing the absolute MSLP values associated with each weather pattern, with the centres of high and low pressure also indicated. Next to the WP labels are the annual (A), winter (W; DJF) and summer (S; JJA) relative frequencies of occurrences of each WP (%). The frequencies of occurrence data are associated with the WPs based on ERA-Interim between 1979 and 2017, while the WP definitions were generated from a clustering process applied to EMULATE MSLP reanalysis data between 1850 and 2003. See the text for details.

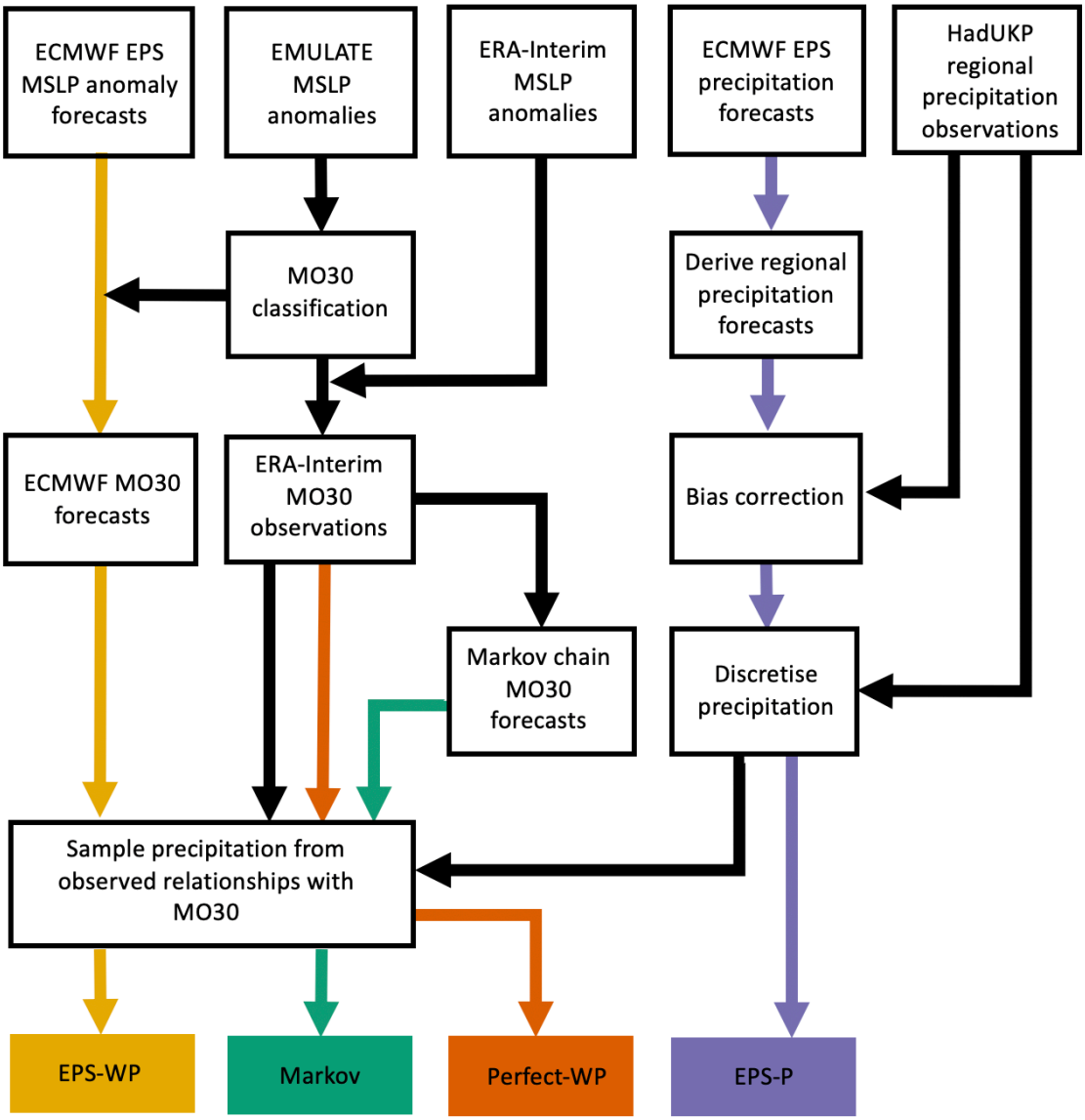

Figure 2: Schematic showing the procedure for the four precipitation forecast models. The top row shows the base data sets used and the bottom row shows the four models. Coloured arrows begin at the first stage for which forecasts are issued: EPS-WP forecasts begin with the ECMWF prediction system MSLP forecasts; Markov forecasts are produced once the ERA-Interim MO30 time series has been derived; Perfect-WP 'forecasts' are observations from the same time series, while EPS-P forecasts are the post-processed data from the ECMWF forecast system.

Figure 3: Jensen-Shannon Divergence scores for EPS-WP and Markov models for three lead-times.

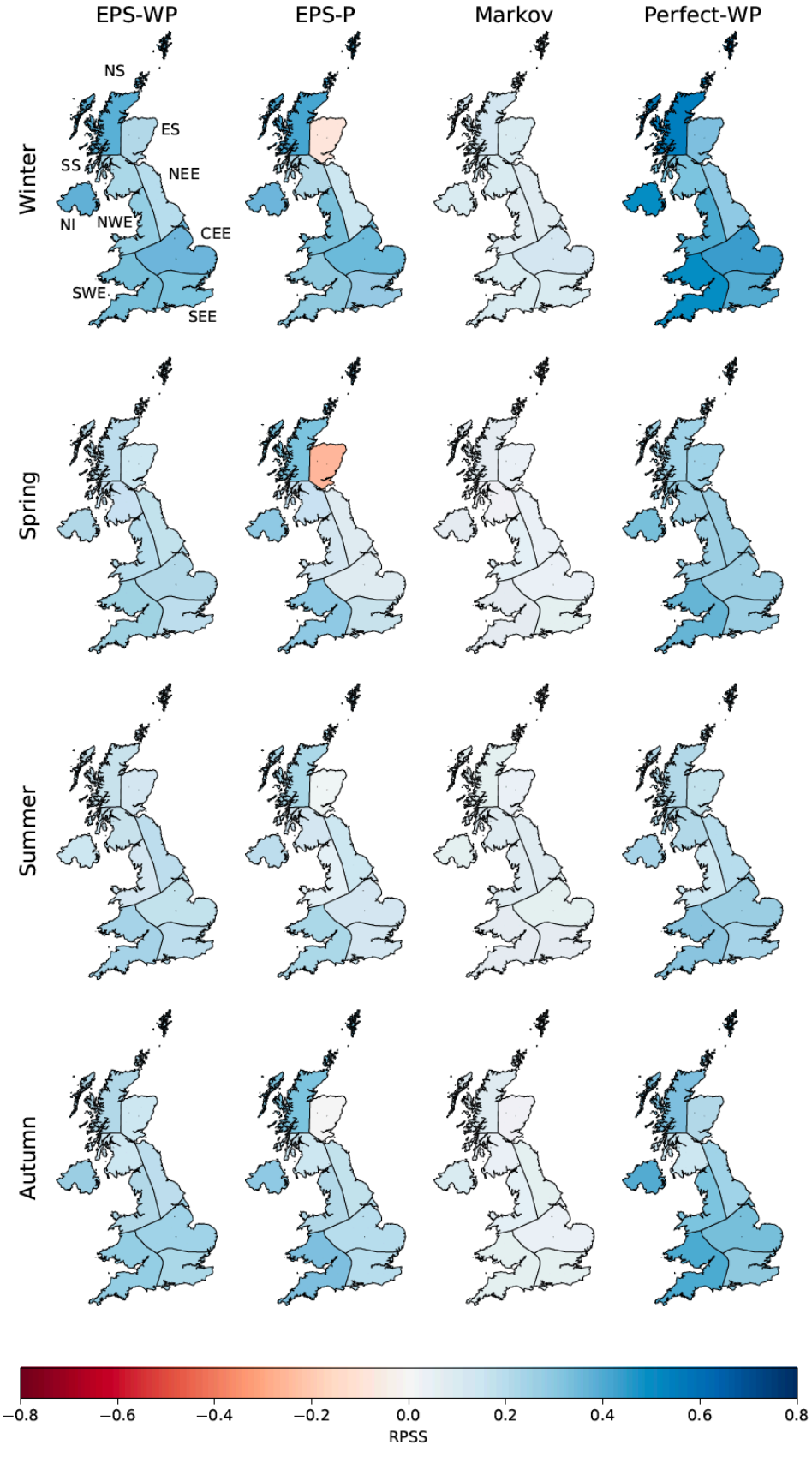

Figure 4: Ranked probability skill scores (RPSS) for precipitation forecasts at a 16-day lead for each model and season.

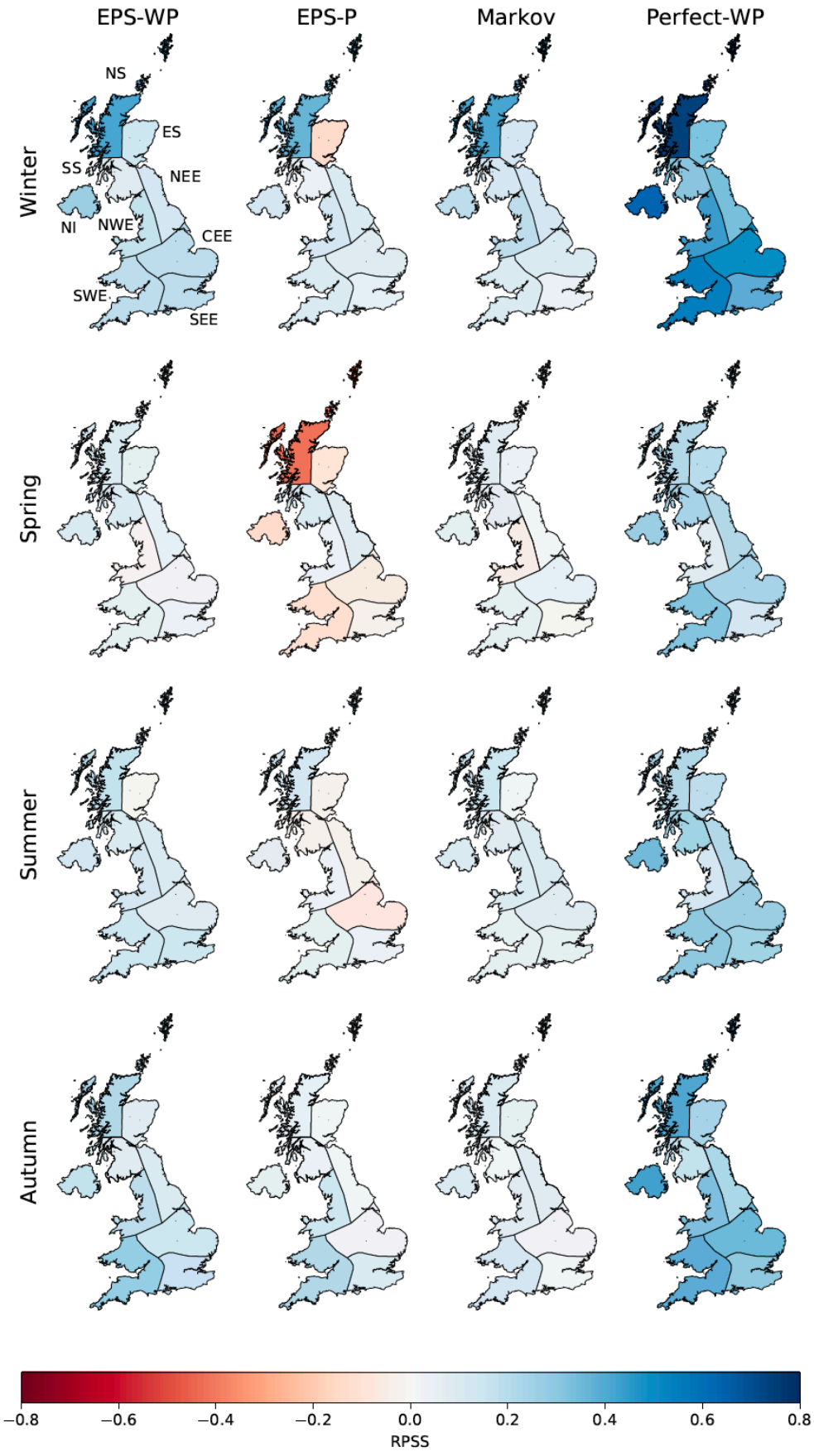

Figure 5: As Figure 4 but for a 46-day lead.

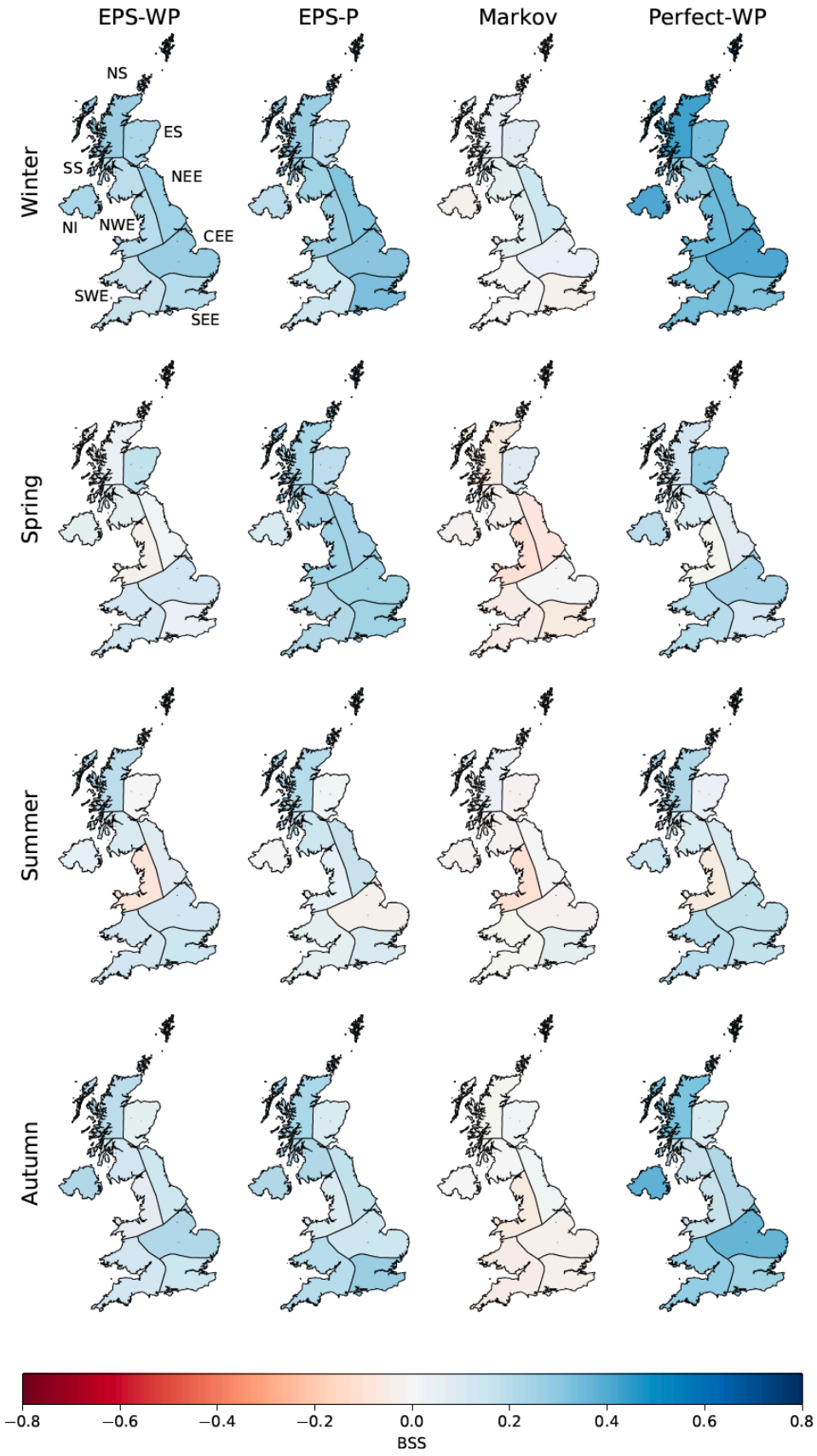

Figure 6: Brier skill scores (BSS) for mild drought (total precipitation below the 30.9th percentile) for a 16-day lead-time for each model and season.

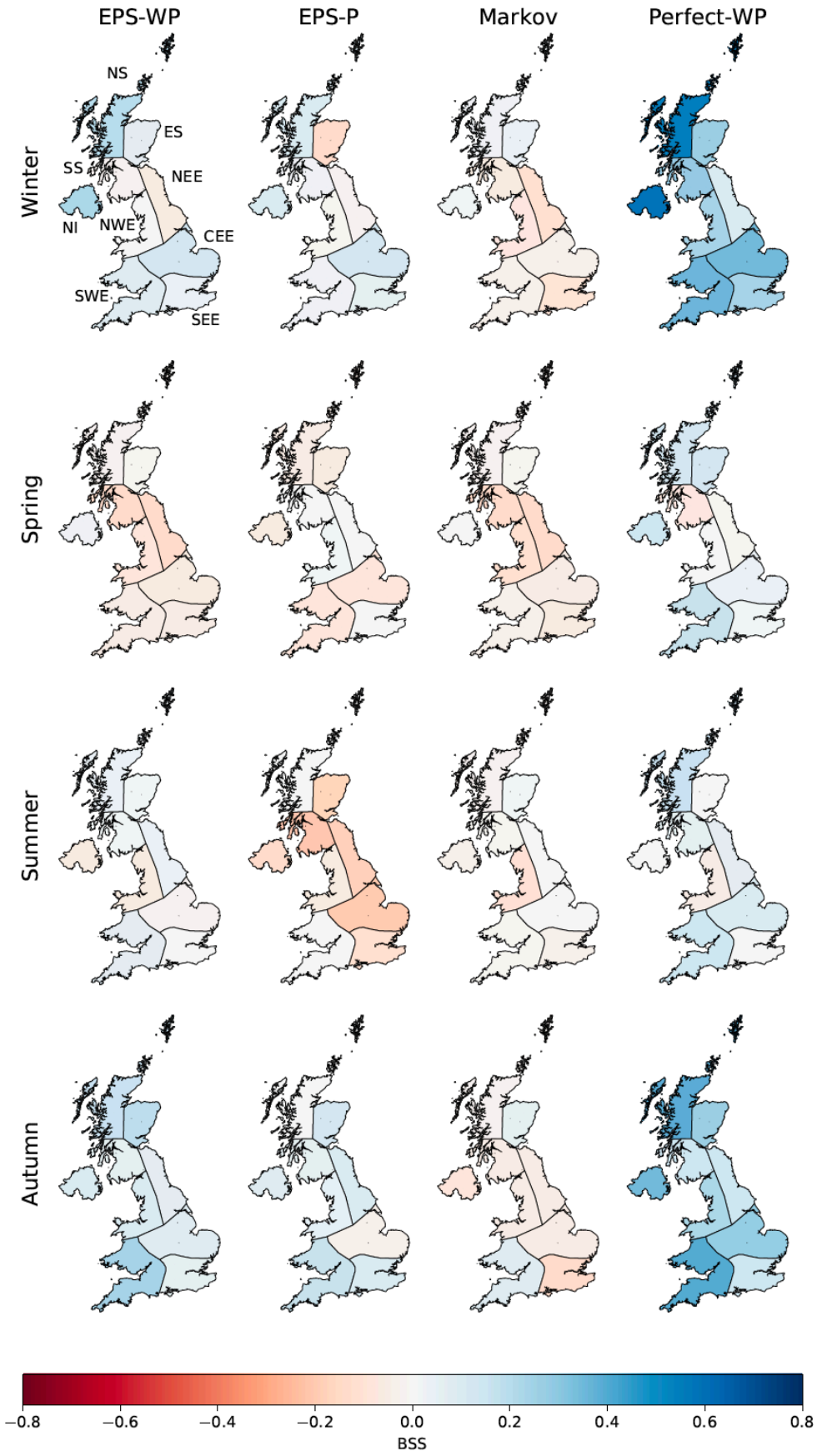

Figure 7: As Figure 6 but for a 46-day lead.

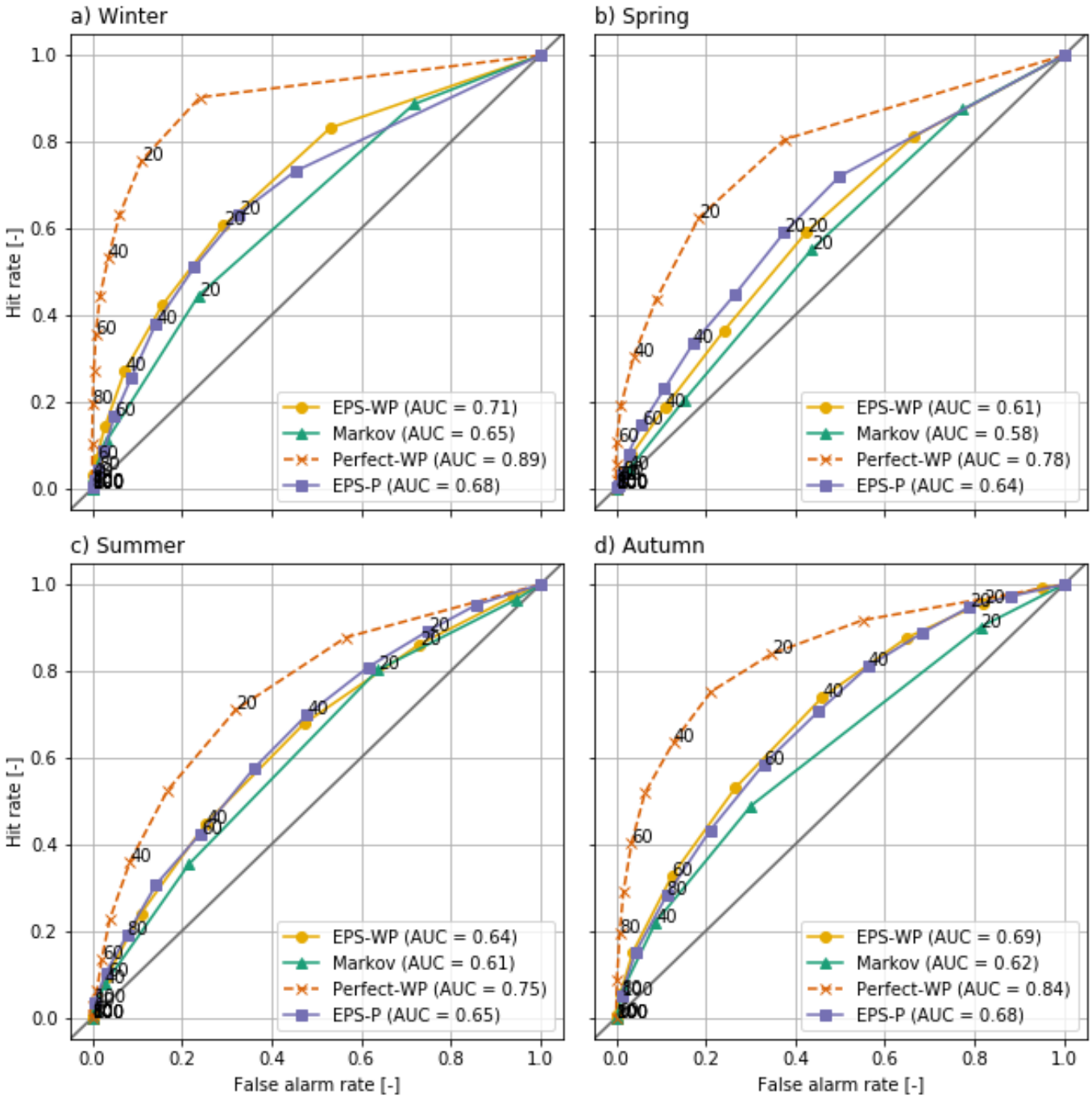

Figure 8: Relative operating characteristics (ROC) curves and area under ROC curve (AUC) for mild drought with a 46-day lead-time. Annotated values indicate drought forecast probability thresholds.

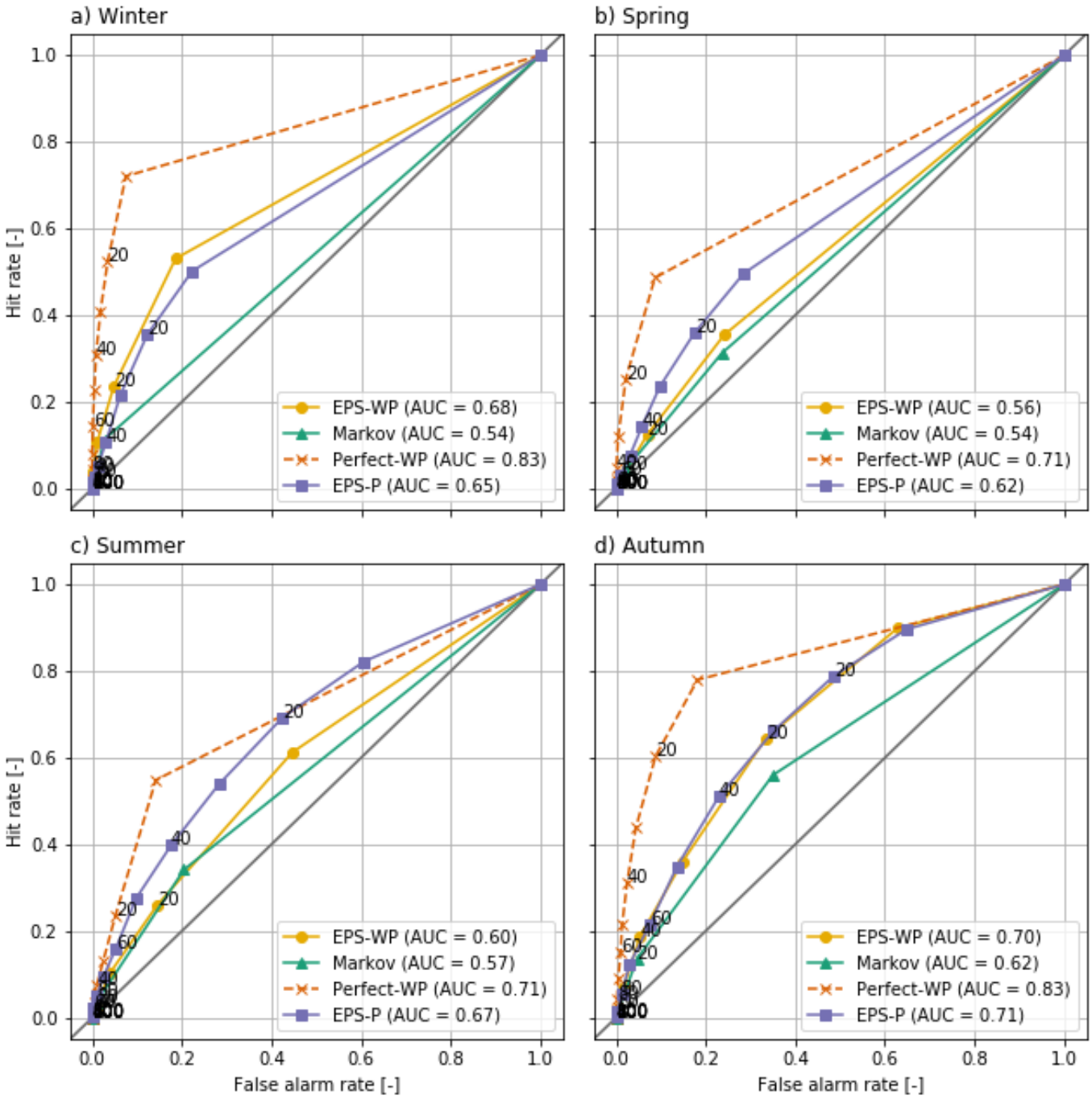

Figure 9: As Fig. 8 but for moderate drought.

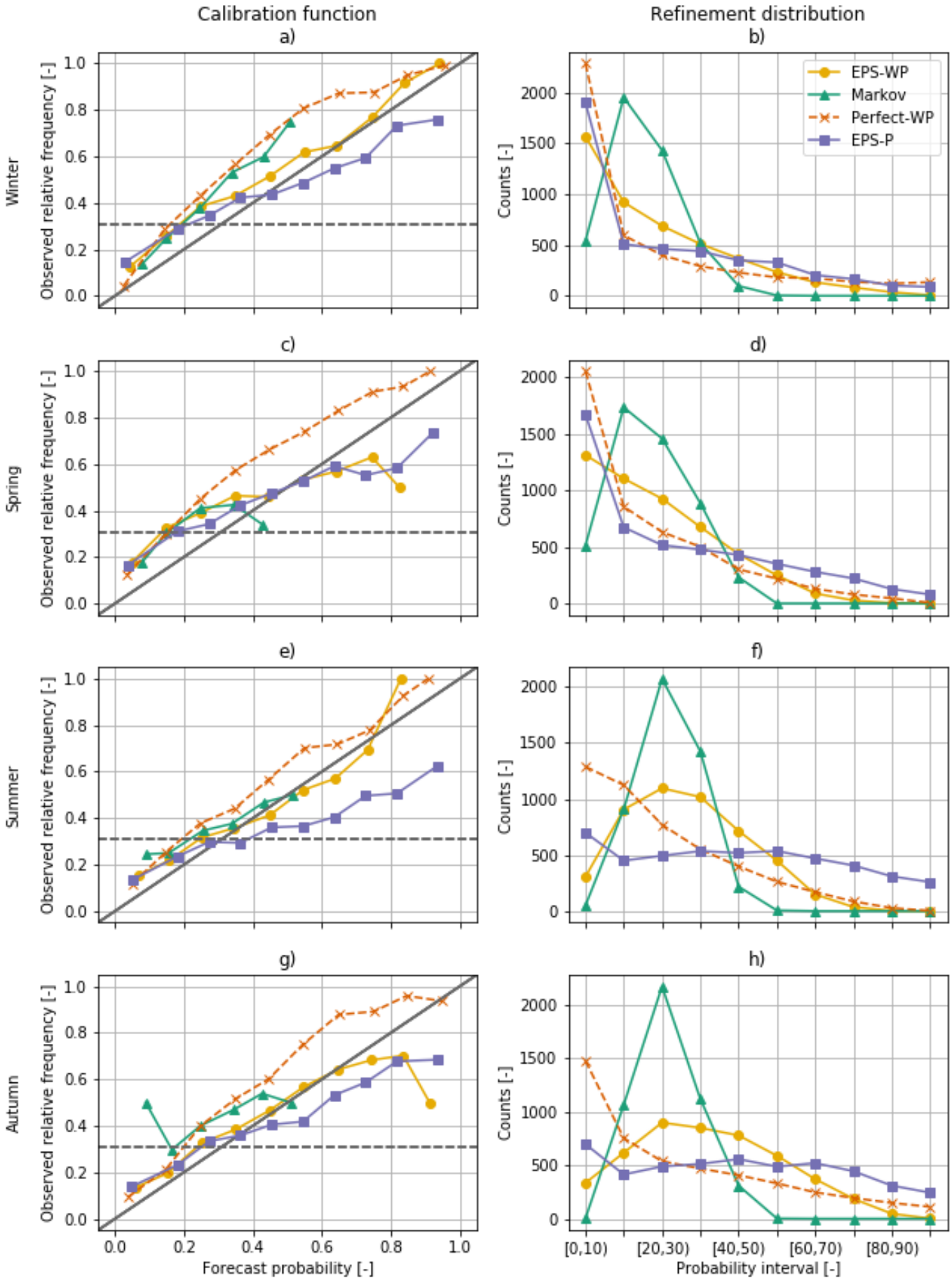

Figure 10: Calibration functions (first column) and refinement distributions (second column) for mild drought with a 31-day lead-time. For the calibration function diagrams, the solid diagonal line indicates perfect reliability and the dashed horizontal line the event relative frequency for mild drought (0.309).

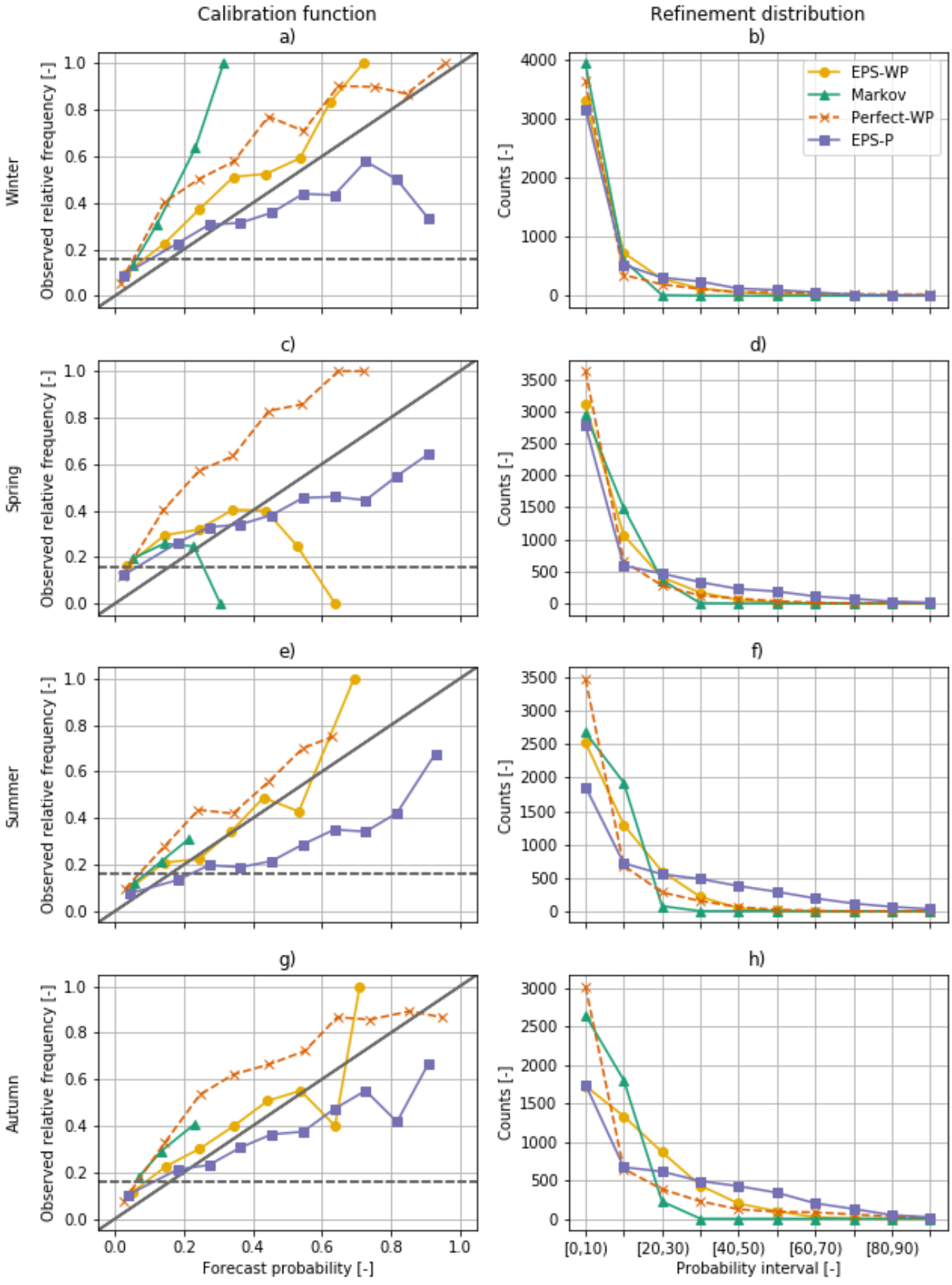

Figure 11: As Fig. 10 but for moderate drought (event relative frequency of 0.159).