Conditioning Ensemble Streamflow Prediction with the North Atlantic Oscillation improves skill at longer lead times

Seán Donegan[1], Conor Murphy[1], Shaun Harrigan[2], Ciaran Broderick[3], Dáire Foran Quinn[1], Saeed Golian[1], Jeff Knight[4], Tom Matthews[5], Christel Prudhomme[2,5,6], Adam A. Scaife[4,7], Nicky Stringer[4], and Robert L. Wilby[5]

[1]Irish Climate Analysis and Research UnitS (ICARUS), Department of Geography, Maynooth University, Co. Kildare, Ireland
[2]Forecast Department, European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK
[3]Flood Forecasting Division, Met Éireann, Dublin 9, Ireland
[4]Met Office Hadley Centre, Exeter, UK
[5]Department of Geography and Environment, Loughborough University, Loughborough, UK
[6]UK Centre for Ecology & Hydrology (UKCEH), Wallingford, UK
[7]College of Engineering, Mathematics, and Physical Sciences, University of Exeter, Exeter, UK

Correspondence: Seán Donegan (sean.donegan@mu.ie)

Abstract. Skilful hydrological forecasts can benefit decision-making in water resources management and other water-related sectors that require long-term planning. In Ireland, no such service exists to deliver forecasts at the catchment scale. In order to understand the potential for hydrological forecasting in Ireland, we benchmark the skill of Ensemble Streamflow Prediction (ESP) for a diverse sample of 46 catchments using the GR4J hydrological model. Skill is evaluated within a 52-year hindcast study design over lead times of 1 day to 12 months for each of 12 initialisation months, January to December. Our results show that ESP is skilful against a probabilistic climatology benchmark in the majority of catchments up to several months ahead. However, the level of skill was strongly dependent on lead time, initialisation month, and individual catchment location and storage properties. Mean ESP skill was found to decay rapidly as a function of lead time, with continuous ranked probability skill scores (CRPSS) of 0.8 (1-day), 0.32 (2-week), 0.18 (1-month), 0.05 (3-month), and 0.01 (12-month). Forecasts were generally more skilful when initialised in summer than other seasons. A strong correlation ($\rho = 0.94$) was observed between forecast skill and catchment storage capacity (baseflow index), with the most skilful regions, the Midlands and East, being those where slowly responding, high storage catchments are located. Forecast reliability and discrimination were also assessed with respect to low and high flow events. In addition to our benchmarking experiment, we conditioned ESP with the winter North Atlantic Oscillation (NAO) using adjusted hindcasts from the Met Office's Global Seasonal Forecasting System version 5. We found gains in winter forecast skill (CRPSS) of 7–18% were possible over lead times of 1 to 3 months, and that improved reliability and discrimination make NAO-conditioned ESP particularly effective at forecasting dry winters, a critical season for water resources management. We conclude that ESP is skilful in a number of different contexts and thus should be operationalised in Ireland given its potential benefits for water managers and other stakeholders.

1 Introduction

Skilful hydrological forecasts at lead times of weeks to months can benefit water resources management (Anghileri et al., 2016; Dixon and Wilby, 2019; Viel et al., 2016; Wetterhall and Di Giuseppe, 2018) and help mitigate extreme events by enhancing preparedness and improving operational decisions (Luo and Wood, 2007; Neumann et al., 2018; Pappenberger et al., 2015a; Zhao and Zhao, 2014). For example, hydrological forecasts have been used to modify reservoir operations for hydropower production (Fan et al., 2016), storage and supply (Turner et al., 2017), and the management of flood and drought conditions (Amnatsan et al., 2018; Ficchì et al., 2016; Watts et al., 2012). They have also been shown to benefit sectors such as agriculture (Mushtaq et al., 2012), tourism (Fundel et al., 2013), and navigation (Meißner et al., 2017). Such applications can yield significant economic returns. For instance, Hamlet et al. (2002) reported a potential rise in annual revenue of $153 million when forecast information was incorporated into the operation of major hydropower dams in the Columbia River basin. Similarly, Pappenberger et al. (2015a) claim that the European Flood Awareness System (EFAS; Thielen et al., 2009) saves around 400 Euro for every 1 Euro invested.

The value of hydrological forecasting has led several countries to establish operational seasonal hydrological forecasting (SHF) systems. These include the U.S. National Weather Service's (NWS) Hydrologic Ensemble Forecast Service (HEFS; Demargne et al., 2014), the Hydrological Outlook UK (HOUK; Prudhomme et al., 2017), and the Australian Bureau of Meteorology's statistical and dynamical forecasts (Schepen and Wang, 2015). Although Ireland benefits from regional hydrological outlooks provided by EFAS, no service currently exists for delivering forecasts at the catchment scale; yet water managers and other stakeholders require confident, locally-tailored forecast information. A national operational SHF system could bridge this gap. However, despite interest from water managers, it is difficult to justify the implementation of such a system as little preparatory work has been done to evaluate the potential for hydrological forecasting in an Irish context.

Recent international assessments of progress in SHF (Tang et al., 2016; Yuan et al., 2015) indicate that: (i) advances in empirical and dynamical SHF are feasible in climate contexts that resemble Ireland; and (ii) SHF spans a wide range of methods with varying complexity and data requirements, but no universally accepted 'best' approach has emerged. As the performance of different methods will likely depend on time of year, lead time, and, critically, local hydrological context (Girons Lopez et al., 2021; Harrigan et al., 2018; Meißner et al., 2017; Pechlivanidis et al., 2020), understanding how best to apply the range of available tools to develop skilful forecasts for Ireland requires rigorous testing at the catchment scale. To the authors' knowledge, only Foran Quinn et al. (2021) have previously evaluated seasonal streamflow forecasts for Ireland. They found that whilst skill was mainly restricted to summer months, statistical persistence forecasts could have practical value in the management of water resources and hydrological extremes. We build on this work and further assess the scientific basis for SHF in Ireland by evaluating and benchmarking the skill of Ensemble Streamflow Prediction (ESP).

ESP is a well-established forecasting technique in which historical sequences of climate data at the time of forecast are used to drive a hydrological model, producing an ensemble of equiprobable future streamflow traces (Day, 1985; Twedt et al., 1977). It is comparable to persistence in that it requires no information about future meteorological conditions; outlooks are instead based on knowledge of hydrological state variables (i.e., antecedent soil moisture, groundwater, snowpack, and streamflow

itself) which can provide predictability up to 5 months ahead (Wood and Lettenmaier, 2008). In this regard, ESP can be used to efficiently specify not only the catchments where knowledge of initial conditions or meteorological forcing may be the greatest source of skill, but also the time of year and lead times over which different skill sources may be dominant (Wood and Lettenmaier, 2006).

The ESP method was originally developed in the snow-dominated catchments of the western United States (e.g., Franz et al., 2003), but has shown skill in other regions, including the UK (Harrigan et al., 2018), European Alps (Förster et al., 2018), Sweden (Girons Lopez et al., 2021), New Zealand (Singh, 2016), Australia (Pagano et al., 2010; Wang et al., 2011), and China (Yuan et al., 2016). Simplicity and efficiency make ESP a popular choice for operational forecasting. It is one of three methods used in the HOUK (Prudhomme et al., 2017) and forms the basis of the NWS HEFS (Demargne et al., 2014). Moreover, ESP is recognised as a low-cost, 'tough-to-beat' forecast (Pappenberger et al., 2015b) against which value-added by more sophisticated hydrometeorological ensemble systems can be assessed (e.g., Arnal et al., 2018; Bazile et al., 2017; Wanders et al., 2019). Hence, the potential application of ESP in Ireland merits exploration.

However, lack of sensitivity to concurrent meteorological conditions limits the application of ESP in areas that are less dependent on the initial hydrological state. Given that local meteorological conditions are known to be teleconnected to regional variations in atmospheric–oceanic modes, ESP techniques may be improved by conditioning on these circulation patterns. Several studies have already demonstrated the added value of incorporating climate information into ESP forecasts in this way. For example, Hamlet and Lettenmaier (1999) found that conditioning ESP traces according to El Niño–Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) indicators significantly improved forecast specificity and extended lead time by about six months in the Columbia River basin. Similarly, both Werner et al. (2004) and Bradley et al. (2015) reported improvements of 28% and 27% in forecast skill, respectively, when conditioning ESP with ENSO. More modest improvements of 5–10% were observed by Beckers et al. (2016) for two test stations when applying an ENSO-conditioned ESP. More recently, Yuan and Zhu (2018) showed that decadal predictions of terrestrial water storage made using ESP could be improved by conditioning with PDO and Atlantic Multidecadal Oscillation indices.

In Europe, the dominant mode of climate variability is the North Atlantic Oscillation (NAO). The NAO affects streamflow predictability, particularly during winter (Bierkens and van Beek, 2009; Steirou et al., 2017; Wedgbrow et al., 2002; Wilby, 2001), and it is highly correlated with winter streamflow over Ireland (Murphy et al., 2013). As winter is the most important season for groundwater recharge in Europe, the ability to accurately forecast winter streamflow would be extremely beneficial for water managers. Advances in predicting the NAO (Scaife et al., 2014; Smith et al., 2020) enable long-range forecasts of UK winter hydrology (Svensson et al., 2015) as well as improved seasonal meteorological forecasts for driving hydrological models (Stringer et al., 2020). Hence, it may be possible to leverage this predictability to improve ESP performance by sub-sampling ensemble members for Ireland using the winter NAO.

In this paper, we benchmark ESP skill against streamflow climatology within a 52-year hindcast study design. Skill is evaluated for a combination of different lead times and initialisation months, and for diverse hydroclimate regions and catchment types. The relationship between catchment characteristics and ESP skill is explored. Reliability and discrimination are assessed

with respect to low and high flow events. We also examine the effect of conditionally sampling ensemble members on ESP skill during winter. The following research questions are addressed:

1. When is ESP skilful, given a wide range of lead times and initialisation months?
2. Where is ESP most skilful, at regional and catchment scales?
3. How does ESP skill relate to catchment characteristics?
4. To what extent can winter ESP skill be improved by conditioning on the NAO?
5. What is the potential for operationalising the ESP method for hydrological forecasting in Ireland?

Section 2 describes our data and methods. Our results are presented in Sect. 3. We offer discussion and suggestions for future research in Sect. 4. Conclusions are presented in Sect. 5.

2 Data and methods

2.1 Catchment selection and observed data

Forty-six catchments were selected for our analysis following the same criteria used to establish the Irish Reference Network (Murphy et al., 2013). Catchments were selected provided they: (i) had quality-assured, long-term observational data, with a minimum record length of 25 years; (ii) had a flow regime which had not been significantly altered by human activity; (iii) had little evidence of land-use change; and (iv) together build a representative sample of Ireland's diverse hydrological and climatological conditions, with good spatial coverage. This selection process ensured sufficient data for hydrological model calibration whilst limiting the potential for confounding factors that could adversely affect the interpretation of results. Catchments were grouped according to the European Union's NUTS (Nomenclature of Territorial Units for Statistics) III regions (Fig. 1) to explore spatial variations in skill. As the Dublin region contained only one catchment in our sample, this was merged with the Mid-East into a single region: The East. The distribution of catchments within the seven regions ranges from four in the West to 10 in the Mid-West. Although the NUTS III regions do not inherently lend themselves to hydrological analysis, grouping the catchments in this way did yield regions that were diverse in terms of their hydrology and climate. They are therefore suitable to examine how skill may differ between areas with contrasting hydroclimate properties.

Observed daily mean streamflow data ($\mathrm{m^3\,s^{-1}}$) were obtained from gauging stations administered by the Office of Public Works (OPW) and the Environmental Protection Agency. Despite the strict selection criteria, some catchments still contain multiple or extended periods of missing data. Hence, streamflow records were retrieved only for calendar years 1992–2017 – the longest usable period common to all 46 catchments. Catchment average daily precipitation ($\mathrm{mm\,d^{-1}}$) and temperature ($^\circ\mathrm{C}$) spanning 1961–2017 were derived from gridded ($1\,\mathrm{km} \times 1\,\mathrm{km}$) datasets developed by Met Éireann (Walsh, 2012). Potential evaporation ($\mathrm{mm\,d^{-1}}$) was calculated from temperature and radiation according to Oudin et al. (2005).

Data on catchment physical attributes were based on a selection of physical catchment descriptors (PCDs) from the OPW's Flood Studies Update (Mills et al., 2014). These PCDs describe facets of catchment hydrology, morphology, soil, and climate, and are used here to examine relationships between catchment characteristics and ESP skill. The primary PCDs of interest are the baseflow index (BFI), the Richards–Baker flashiness index (RBI; Baker et al., 2004), and the runoff ratio (RR), as these describe aspects of catchment storage and response and have been linked to ESP skill (e.g., Girons Lopez et al., 2021; Harrigan et al., 2018; Pechlivanidis et al., 2020). The BFI is calculated according to the Institute of Hydrology method (Gustard et al., 1992) and quantifies the contribution of stored sources to runoff. Hence, the BFI can be considered an integrated measure of catchment storage capacity. The RBI measures the frequency and rapidity of short-term changes in streamflow, and the RR gives the amount of runoff relative to the amount of precipitation received. Across the sample of catchments, median (5^{th} and 95^{th} percentile) BFI is 0.59 (0.34, 0.75), median RBI is 0.19 (0.07, 0.5), and median RR is 0.62 (0.5, 0.82). Higher values of RBI and RR are observed for catchments with lower storage capacity (BFI) and smaller area, indicative of more responsive hydrological regimes. In addition to the BFI, we also represent catchment storage using the calibrated GR4J x_1 and x_3 parameters, the sum of which give an overall indicator of storage capacity. Catchment area ranges from $5.46\,\mathrm{km}^2$ to $2460\,\mathrm{km}^2$. Although snow has been shown to be a major source of hydrological predictability (e.g., Greuell et al., 2019; Shukla et al., 2013; Wood and Lettenmaier, 2008), it is not known to make a substantial contribution to precipitation in Ireland. No catchments have a significant amount of snowfall, defined following Berghuijs et al. (2014) as a long-term mean fraction of precipitation falling as snow ($\overline{F_s}$) < 0.15. Hence, we do not consider the role of snow in our analysis. A complete list of PCDs referred to in this study is given in Table 1. Catchment characteristics are summarised for Ireland and each of the NUTS III regions in Table 2, and for individual catchments in Table S1 in the Supplement.

2.2 Hydrological modelling

The GR4J (Génie Rural à 4 paramètres Journalier; Perrin et al., 2003) daily lumped conceptual rainfall-runoff model was applied. This model has a parsimonious structure consisting of four free parameters (x_1–x_4) that require calibration of observed streamflow data against precipitation and potential evaporation. The model structure can be described in terms of its water balance and routing operators (Santos et al., 2018). Water is partitioned between a production (soil moisture accounting) store and a routing store. The production store (capacity $x_1\ \mathrm{mm}$) gains water from rainfall and loses water from evaporation and percolation. Ninety percent of the total quantity of water reaching the routing component (i.e., the sum of the percolation leak and the water bypassing the production store) is routed by a single unit hydrograph (time base $x_4\ \mathrm{d}$) and a non-linear routing store (capacity $x_3\ \mathrm{mm}$). The remaining 10% is routed by a single unit hydrograph (time base $2(x_4)\ \mathrm{d}$). A groundwater exchange function (rate $x_2\ \mathrm{mm\,d^{-1}}$) operates on both routing channels and can be positive, negative, or zero.

We chose GR4J on the basis of its reliability. The model has undergone extensive testing in several countries and has been shown to accurately simulate the hydrology of diverse catchment types, with comparatively good results (e.g., Coron et al., 2012; Perrin et al., 2003; Vaze et al., 2011). It has also been successfully applied to Irish conditions (Broderick et al., 2016, 2019) where it was found to perform well for a similar set of catchments to those used here, with respect to both temporal transition between contrasting climate periods and the reproduction of various hydrological signatures. Moreover, GR4J has

been used previously for ESP (Harrigan et al., 2018; Pagano et al., 2010). We find the model uniquely suited to this application, as large ensembles of runs are required in long hindcast experiments. These simulations can be computationally intensive and time consuming with more complex model structures, which do not necessarily lead to large improvements in skill (e.g., Bell et al., 2017). GR4J is implemented in R via the open-source 'airGR' package (v1.4.3.65; Coron et al., 2017, 2020).

Model parameters were estimated using Memetic Algorithms with Local Search Chains (MA-LS-Chains; Bergmeir et al., 2016; Molina et al., 2010). As ESP forecasts are made throughout the year under varying conditions, the non-parametric Kling–Gupta efficiency (KGE_{NP}; Appendix A) was chosen as the objective function to optimise, as it has been shown to capture multiple parts of the hydrograph well (Pool et al., 2018). Parameter estimation was carried out in R using the 'Rmalschains' package (v0.2-6; Bergmeir et al., 2016, 2019) with the Covariance Matrix Adaptation Evolution Strategy (Hansen and Ostermeier, 2001) as the local search method.

Model calibration was performed following the procedures recommended by Arsenault et al. (2018). A split-sample test (Klemeš, 1986) was first used to assess model robustness. The available record was divided into two periods of equal length, denoted here as period 1 (P1; 1 January 1993–2 July 2005) and period 2 (P2; 2 July 2005–31 December 2017). Separate parameter sets were created using data from P1 and P2 in turn for calibration and validation (i.e., parameters were calibrated on P1 and validated on P2 and vice versa). A third round of calibration was then performed using data from the complete period (CP; 1 January 1993–31 December 2017). This parameter set was carried forward for all subsequent modelling tasks. An approach of this nature is beneficial as it allows for evaluation of the model's ability to accurately simulate catchment processes over two independent periods whilst maximising the information content of the parameter set that is used to generate the ESP hindcast time series. In all cases, 1992 was used as a warm-up period to initialise model states, and the full series (1993–2017) was simulated before calibration and testing to preserve the internal dynamics and temporal stability of catchment stores. Model performance was evaluated using KGE_{NP}, the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970), and the percent bias (PBIAS; Gupta et al., 1999).

2.3 ESP study design

2.3.1 Historical ESP

Forecasts were initialised on the first day of each month following a 4-year model warm-up period to estimate initial hydrological conditions. The first usable forecast date after model warm-up is, therefore, 1 January 1965. For each forecast initialisation date, a 55-member ensemble *m* of streamflow hindcasts was generated by forcing GR4J with corresponding historic climate sequences (pairs of precipitation and potential evaporation) extracted from 1961–2016 out to a 12-month lead time. Following Harrigan et al. (2018), streamflow at a given lead time is expressed as the mean daily streamflow from the forecast initialisation date to *n* days or months ahead in time. For example, a January forecast with a lead time of 1 month is the mean daily streamflow from 1 January to 31 January, and a January forecast with a lead time of 2 months is the mean daily streamflow from 1 January to 28 February. Average flow values are used, particularly at monthly time scales, because these are preferred by decision-makers in many water sectors (Arnal et al., 2018). Hindcast time series were therefore temporally aggregated to

provide predictions of mean streamflow over lead times of 1 day to 12 months, resulting in 365 lead times per forecast (excluding leap days). In order to mimic operational conditions and prevent artificial skill inflation (see Robertson et al., 2016), we also employed leave-one-out cross-validation (L1OCV) whereby data from the forecast year was not used as input to the model, as this would not be available in a real-time forecasting setting. For example, a forecast initialised on 1 January 1965 will use historic climate sequences of 365 days in length (1 January to 31 December) extracted from 1961–2016, but not 1965. ESP skill is evaluated over 52 initialisation years N (1965–2016) with 12 initialisation months i (January to December). In total, 624 hindcasts were generated ($N \times i$) with 34,320 individual ensemble members ($N \times i \times m$), each at 365 lead times across 46 catchments, resulting in a hindcast archive of more than 5.7×10^8 streamflow values.

2.3.2 Conditioned ESP

To investigate the potential for improving winter streamflow predictability, we conditioned the ESP method using adjusted NAO hindcasts from the Met Office's Global Seasonal Forecasting System version 5 (GloSea5; MacLachlan et al., 2015). GloSea5 is built around the high-resolution Hadley Centre Global Environmental Model version 3 (HadGEM3) which integrates atmosphere, ocean, land, and sea-ice components. HadGEM3 has an atmospheric resolution of 0.83° longitude by 0.55° latitude with 85 vertical levels and an ocean resolution of 0.25° in both latitude and longitude with 75 vertical levels. Although GloSea5 has been shown to skilfully predict the NAO (Scaife et al., 2014), several studies have documented a signal-to-noise problem that limits the usefulness of forecasts to drive hydrological models, as ensemble mean signals in NAO forecasts are anomalously weak (Eade et al., 2014; Scaife et al., 2014; Scaife and Smith, 2018). Focusing on the dynamical signals can correct this by amplifying the ensemble mean (Baker et al., 2018), so adjusted hindcasts are used here following the method of Stringer et al. (2020). For each DJF period over 1993–2016, we combined GloSea5 hindcasts initialised on 1, 9, and 17 November, each with 17 ensemble members, to create a 51-member lagged ensemble of raw NAO predictions. After adjustment to remove the signal-to-noise discrepancy in the raw ensemble, predicted monthly NAO values were used to select 10 non-sequential DJF analogues (e.g., December 2007, January 1980, February 2011) where the mean observed seasonal NAO approximated the mean adjusted seasonal NAO hindcast. This resulted in a 510-member ensemble of analogue date sequences which were then used to extract corresponding precipitation and potential evaporation for input to the ESP method. The decision to construct analogue seasons with months from different years was made to: (a) ensure that the range of possible values suggested by GloSea5 could be reproduced; and (b) to avoid underestimating extreme seasonal NAO values, which would sample exclusively from DJF 2009–10 if below $-10\,\mathrm{hPa}$ (Stringer et al., 2020). Ten analogues were sampled per hindcast member to minimise non-NAO-related variability whilst keeping a consistent NAO signal across the sample. Conditioned ESP forecasts were only initialised on 1 December. A more detailed description of the adjustment procedure and the selection of the analogue date sequences is available in Stringer et al. (2020).

2.4 Skill evaluation

2.4.1 Hindcast overall performance

We quantify the overall skill of the ESP method using the continuous ranked probability score (CRPS; Hersbach, 2000) and corresponding skill score (CRPSS; Appendix B). The CRPS is a recommended and widely-used evaluation metric for ensemble hydrological forecasting (Pappenberger et al., 2015b) that penalises biased and unsharp forecasts (Wilks, 2019). To minimise the impact of hydrological model uncertainty on hindcast quality, we use modelled observations derived from GR4J in place of direct streamflow data when evaluating skill. This is common practice (e.g., Arnal et al., 2018; Harrigan et al., 2018; Wood and Lettenmaier, 2008; Wood et al., 2016) as it isolates loss of skill to errors in initial conditions. Our reference forecast is constructed as the full-sample climatological distribution of modelled observations over 1965–2016 for the forecast period. This forecast was also created using L1OCV to account for streamflow persistence. In the case of the conditioned ESP, skill is calculated relative to both the probabilistic climatology benchmark and the full historical ESP ensemble. In all cases, the Ferro et al. (2008) ensemble size correction for CRPS is applied after cross-validation to account for differences in the number of ensemble members.

2.4.2 Hindcast reliability

Hindcast reliability was also assessed for low and high flows. Reliability refers to the overall agreement between the forecast probabilities and the observed frequencies. For each catchment, initialisation month, and lead time, the probability integral transform (PIT; Gneiting et al., 2007; Laio and Tamea, 2007) score was calculated for subsets of forecast–observation pairs falling within the lower and upper terciles of the corresponding modelled observations. The PIT score was derived from the PIT diagram following Renard et al. (2010). A forecast with a PIT score of 1 has perfect reliability whereas a forecast with a PIT score of 0 has the worst reliability.

2.4.3 Hindcast discrimination

Hindcasts were further assessed in terms of their ability to discriminate between events and non-events using the receiver operator characteristic (ROC; Mason and Graham, 1999) score. The ROC score is defined as the area under the ROC curve, which plots the probability of detection against the probability of false detection for a given event and a range of probability levels (Demargne et al., 2010). A ROC score of 1 indicates that all ensemble members correctly predicted the event in all years, whereas a ROC score of 0.5 indicates a forecast with no discrimination. For each catchment, initialisation month, and lead time, the ROC score was calculated using the lower and upper terciles of the corresponding modelled observations as thresholds. Hence, the ROC score should be interpreted as a measure of how well ESP can forecast the occurrence of low and high flow events and can thus be regarded as an indicator of potential usefulness. We use a slightly stricter skill threshold of 0.6 so that forecasts are only considered skilful if they are better than guesswork. Both the CRPSS and ROC score were calculated in R using the 'easyVerification' package (v0.4.4; MeteoSwiss, 2017).

3 Results

3.1 Hydrological model performance

GR4J performed well for our catchment sample (Fig. 2). The median (5^{th} and 95^{th} percentile) value of KGE_{NP} is 0.95 (0.88, 0.97) for calibration over P1, P2, and CP. Median validation scores of 0.91 (0.84, 0.96) were achieved during testing on both P1 and P2. Median NSE for calibration over CP is 0.88 (0.69, 0.93) and median PBIAS is 0.04% (−0.13%, 0.14%). Performance metrics and calibrated parameter values for individual catchments over CP are given in Table S1.

3.2 Timing of ESP skill

3.2.1 Lead time

Mean ESP skill declines rapidly as a function of lead time, across all catchments and initialisation months (Fig. 3). Mean CRPSS values for short (1-day) to extended (2-week) lead times range from 0.8 to 0.32, and for monthly (1- and 2-month), seasonal (3-month), and annual lead times from 0.18, 0.09, and 0.05, to 0.01, respectively. However, the rate at which skill decays across catchments varies, with considerable differences around the mean shown by the 5^{th} and 95^{th} percentile bands. For example, for a 2-week lead time CRPSS values within this band range between 0.1 and 0.58, and for a 1-month lead time between 0.03 and 0.4.

3.2.2 Initialisation month

ESP skill varies with forecast initialisation month and time of year, with highest and lowest skill scores dependent on lead time (Fig. 4). For short to monthly lead times, skill scores are highest when forecasts are initialised in summer (JJA), with July the most skilful initialisation month on average, whereas skill tends to be lower during winter (DJF), with January and December exhibiting the lowest skill. At seasonal lead times, skill during autumn (SON) is comparable to that of summer, whilst the least skilful forecasts are produced in the spring months (MAM). As in Fig. 3, skill tends toward zero as lead time increases, regardless of initialisation month. Although this decline in performance is less severe for summer than for other seasons, by a 12-month lead time nearly all forecasts are less skilful than climatology. Despite this, several catchments have above (below) average skill scores with some performing notably better (worse) across different lead times and initialisation months. For example, ESP forecasts initialised in July with a 1-month lead have moderate skill on average (CRPSS $= 0.34$), but seven catchments have high skill (CRPSS ≥ 0.5) with a maximum CRPSS of 0.68 for the Erkina (ID 15005). Conversely, 14 catchments have low skill (CRPSS ≤ 0.25) with a minimum of -0.03 for the Newport (ID 32012).

3.3 Spatial distribution of ESP skill

3.3.1 NUTS III regions

Mean ESP skill across all initialisation months is shown in Fig. 5 for Ireland and each of the seven NUTS III regions. The Midlands, Mid-West, and East are the most skilful regions, followed by the South-East, West, and Border regions. The South-West is the least skilful region on average, with the lowest CRPSS values for all sampled lead times. Regional variations in skill are less pronounced at shorter lead times but become more apparent as lead time increases. For example, at a 1-month lead time the Midlands (CRPSS $= 0.26$) is twice as skilful as the Border (CRPSS $= 0.13$) and South-West (CRPSS $= 0.12$). All regions are, on average, skilful out to a 1-month lead time, but the Midlands is the only region that is moderately skilful (CRPSS ≥ 0.25). The Midlands remains the most skilful region beyond 1-month, though the level of skill is generally quite low for all regions by this point. The regional variations observed in Fig. 5 are partly explained by the relationship between catchment characteristics and ESP skill (Sect. 3.4) as the pattern is broadly consistent with differences in catchment storage capacity and wetness. For instance, the Midlands has a high median BFI of 0.71, a low median RBI of 0.13, and a low median SAAR of $939\,\mathrm{mm}$, whereas the South-West has a low median BFI of 0.44, a high median RBI of 0.4, and a high median SAAR of $1407\,\mathrm{mm}$. Differences in regional hydroclimate properties therefore contribute to differences in regional skill as forecasts perform better in the baseflow dominated catchments of the Midlands than the flashy, wetter catchments of the South-West.

3.3.2 Catchment scale

Notable sub-regional heterogeneity emerges when examining skill scores for individual forecasts at the catchment scale (Fig. 6). This heterogeneity is more noticeable at monthly to seasonal lead times, where skilful forecasts are possible for several catchments at different times of the year, even if average skill for the region as a whole tends to be low. For example, whilst the South-West is the least skilful region at a 1-month lead time, with an average CRPSS of 0.12, forecasts with above average skill are possible in several catchments in the region in June, such as the Blackwater (ID 18003; CRPSS $= 0.25$) and the Laune (ID 22035; CRPSS $= 0.23$).

3.4 Relationship with catchment characteristics

Figure 7 shows the relationship between ESP skill, as represented by the average 1-month CRPSS, and several PCDs for each of the 46 study catchments using the non-parametric Spearman rank correlation coefficient (ρ). ESP skill is closely linked with catchment storage properties and responsiveness. There are strong positive correlations between modelled storage capacity (x_1 + x_3) and BFI ($\rho = 0.79$) and between ESP skill and BFI ($\rho = 0.94$). There is also a strong positive correlation between ESP skill and modelled storage capacity ($\rho = 0.75$). Conversely, there is a strong negative correlation between ESP skill and the RBI ($\rho = -0.82$) and a moderate negative correlation between ESP skill and the RR ($\rho = -0.63$). All of these correlations are statistically significant ($p \leq 0.05$). In general, ESP skill tends to be higher for slower responding catchments with greater storage capacity, and lower for faster responding, flashy catchments with poor infiltration. ESP skill is also positively correlated

with catchment area ($\rho = 0.5$) and main-stream length ($\rho = 0.46$), indicating a tendency for the method to perform better in larger catchments with longer streams. Negative correlations exist between ESP skill and PCDs related to catchment wetness (SAAR, FLATWET, and PEAT), though these PCDs also exhibit negative correlations with BFI and positive correlations with RBI and RR, highlighting that wetter catchments are more likely to be those with lower storage and flashier regimes in which ESP has already been shown to perform poorly. Poor skill in these catchments is likely a combination of high precipitation and low permeability which leads to more variable hydrological conditions as rainfall events propagate to streamflow quickly. Finally, there are moderate negative correlations between ESP skill and S1085 ($\rho = -0.67$) and TAYSLO ($\rho = -0.59$) indicating that forecasts are less skilful in catchments with steeper gradients. Although these results are based on the 1-month CRPSS averaged across all initialisation months, similar results are observed for a variety of different months and lead times (not shown).

3.5 Reliability of low and high flow forecasts

ESP is capable of producing reliable forecasts of both low (lower tercile) and high (upper tercile) flows (Fig. 8). However, the level of reliability is dependent on both lead time and initialisation month. Reliability decreases as lead time increases, though the rate at which this occurs is not uniform across all initialisation months. Furthermore, there is considerable inter-catchment variability for both low and high flow forecasts. This latter point is perhaps most pronounced at short to extended lead times, but is also evident at longer leads (e.g., 1- and 2-month forecasts initialised in June and July) where some catchments return much higher than average PIT scores. Reliability tends to be highest when forecasts are initialised in summer and lowest when initialised in winter, with the smallest and largest reductions in PIT scores also evident for these seasons as lead time increases. Across all lead times and initialisation months, reliability is, on average, higher for low flow forecasts than high flow forecasts. Although the PIT score decays with lead time, unlike the CRPSS it does not tend toward zero and instead has a lower bound of around 0.3. Hence, somewhat reliable forecasts of both low and high flows are still possible at annual lead times even when overall skill (CRPSS) is poor.

3.6 Discrimination between events and non-events

In general, ESP is skilful at forecasting the occurrence of both low (lower tercile) and high (upper tercile) flow events up to 1 month ahead in the majority of catchments and for all initialisation months (Fig. 9). Discrimination for both event types is also possible at lead times of 2 and 3 months, though to a lesser extent. These results highlight that ESP still has utility at longer lead times, even when overall performance as measured by the CRPSS is poor. However, this utility seldom extends beyond 3 months, except for specific catchments and initialisation dates, with little or no skill at lead times of 6 and 12 months across the majority of the catchment sample. Some seasonality in ROC skill is apparent, particularly at monthly lead times, where ESP can more skilfully discriminate between events and non-events in summer than other seasons. Discrimination is more skilful for low flow events than high flow events.

3.7 Improvements in winter skill

The overall skill (CRPSS) of NAO-conditioned ESP is compared with that of historical ESP in Fig. 10. Whilst historical ESP is skilful in the majority of catchments at a 1-month lead time, there is a dramatic reduction in both the magnitude of skill and the number of catchments for which skilful forecasts can be made at 2- and 3-month lead times. NAO-conditioned ESP outperforms historical ESP relative to the climatology benchmark in all but one catchment at a 1-month lead time, though these improvements are generally modest, with a median (5^{th} and 95^{th} percentile) difference in CRPSS of 0.04 (0.01, 0.07). At a lead time of two months, NAO-conditioned ESP remains skilful against climatology in 98% of catchments, compared to historical ESP which is only skilful in 37% of catchments. The value of the NAO-conditioned ESP is more evident at a 3-month lead time, where skilful forecasts are still possible for several catchments in the Border and western regions, when historical ESP exhibits little or no skill across the majority of the sample.

Over the three lead times examined here, the greatest improvements are found for wet, fast responding catchments with low baseflow contribution. For example, two of the best performing catchments for NAO-conditioned ESP are the Owenea (ID 38001) and the Fern (ID 39009). The Owenea has a BFI of 0.27, the lowest in the sample, with high SAAR ($1753\,\mathrm{mm}$), RR (0.82), and RBI (0.58) values. The Fern has a below-average BFI of 0.47 with similarly high SAAR and RR values of $1570\,\mathrm{mm}$ and 0.79, respectively, although it is not as flashy (RBI $= 0.18$). NAO-conditioned forecasts generally perform the worst in slowly responding catchments with high storage capacity. At a lead time of 3 months, negative skill is observed in several catchments in the East and South-East, though these values can still be defined within the bounds of what Bennett et al. (2017) refer to as 'neutral skill' (± 0.05 CRPSS) and hence do not represent a significant departure from the performance of historical ESP. These differences in performance can be explained by the relative contribution of initial conditions and meteorological forcing to ESP skill. In the flashy catchments where NAO-conditioned ESP performs well, meteorological conditions are the dominant control on skill as rainfall events propagate to streamflow at a faster rate and memory of initial conditions is lost quickly. It is also worth noting that in these catchments skill generally increases with lead time. This is likely due to the fact that the underlying NAO signal is not as strong over shorter averaging periods due to the noise of the individual weather systems. Moreover, only the seasonal mean NAO is re-scaled to account for the signal-to-noise problem when adjusting hindcasts, so skill is only present at the longer 3-month lead time. For example, at a 3-month lead time NAO-conditioned ESP improves forecast skill by $\sim$18% over historical ESP in both the Owenea and Fern, whereas gains of 7% and 12% are observed for 1- and 2-month lead times, respectively. Conversely, catchments where negative skill is observed have high baseflow contribution and long recession times. Hence, hydrological response is controlled predominately by the slow release of water from reservoirs and initial conditions act as the primary source of skill. The combination of initial conditions and subsampled climate information grant modest improvements in skill in these catchments up to a 1-month lead time. However, at longer lead times improved atmospheric representation alone cannot compensate for divergences from the initial state. Skill deteriorates as a result, eventually becoming negative.

In addition to the CRPSS, both the PIT score and the ROC score were calculated for NAO-conditioned ESP. Figure 11 shows the difference between PIT scores calculated for historical ESP and NAO-conditioned ESP at lead times of 1, 2, and 3 months.

Conditioning ESP with the NAO increases the reliability of low flow forecasts in all catchments at a 1-month lead time. Some catchments experience a reduction in low flow reliability at a 2-month lead time, whereas at a 3-month lead time low flow reliability is observed to increase in almost all catchments. High flow reliability increases in some catchments at a 1-month lead time but then decreases in almost all catchments at lead times of 2 and 3 months. At these longer lead times, increases in high flow reliability tend to be restricted to flashy catchments (e.g., Owenea) where NAO-conditioned ESP has already been shown to perform well in terms of CRPSS.

ROC scores for individual catchments and the full range of lead times are presented in Fig. 12. On average, NAO-conditioned ESP extends the lead time over which discrimination between events and non-events is possible by 141% for low flows (37 days to 89 days) and 170% for high flows (33 days to 89 days). These are considerable improvements over historical ESP, which failed to meet the skill threshold in most catchments at longer lead times. For example, skilful discrimination of low flow events are possible in 78% of catchments at a 3-month lead time when using NAO-conditioned ESP compared to only 11% of catchments when using historical ESP. This makes NAO-conditioned ESP particularly effective at forecasting dry winters, which can be critical for water resources management. It is worth noting that in many catchments NAO-conditioned ESP can 'lose' skill before later regaining it, with the ROC score falling only marginally below the skill threshold. Although this is also observed for historical ESP, it is less frequent.

Changes in reliability are generally consistent with improvements in skill (CRPSS) and discrimination (ROC). Improved low flow reliability allows NAO-conditioned ESP to better distinguish between low flow events and non-events. The reductions in low flow reliability in some catchments at a 2-month lead time are also consistent with NAO-conditioned ESP 'losing' ROC skill before later regaining it (Fig. 12). Increases in high flow reliability at a 3-month lead time in flashy catchments correspond with the greatest increases in CRPSS from NAO-conditioned ESP. In these catchments, where streamflow variability is greater and the NAO is most influential, improved reliability and sharpness lead to better overall skill at longer lead times.

4 Discussion

4.1 When is ESP skilful?

For short lead times (1–3 days), ESP forecasts are on average highly skilful (CRPSS ≥ 0.5) and for extended lead times (1–2 weeks) moderately skilful (CRPSS ≥ 0.25). Mean ESP skill decays rapidly with lead time. Hence, forecast skill for monthly, seasonal, and annual lead times is on average much lower. This is because ESP relies on the long-term 'memory' of the hydrological system. The cumulative effect of distinct meteorological forcing causes a divergence from the initial state that grows with time. Thus, ESP suffers at longer lead times as there is little or no persistence of initial hydrological conditions. Over longer periods, we find that ESP is most skilful out to a month ahead (CRPSS $= 0.18$) but that some predictability (CRPSS > 0.05) is possible up to 3 months in advance. This rapid decline in forecast skill is consistent with findings from several other benchmarking experiments, including Harrigan et al. (2018) and Girons Lopez et al. (2021), who noted a similar deterioration in ESP skill in the UK and Sweden, respectively. Pechlivanidis et al. (2020) also reported a decline in seasonal streamflow forecasting skill with increasing lead time across Europe. Persistence forecasts, which also rely on hydrological

memory as their main source of skill, have shown comparable results. For example, both Svensson (2016) and Foran Quinn et al. (2021) noted a reduction in the number of usable persistence forecasts in the UK and Ireland, respectively, when moving from a 1-month forecast horizon to a 3-month forecast horizon.

ESP skill is also highly dependent on initialisation month. On average, at short-to-extended lead times (1 day to 2 weeks), ESP is most skilful when initialised in summer and least skilful when initialised in winter. This is again consistent with previous research, with higher predictability during dry seasons for forecasting methods that rely on hydrological memory reported for the UK (Harrigan et al., 2018), Switzerland (Staudinger and Seibert, 2014), China (Yang et al., 2014), and parts of the Amazon Basin (Paiva et al., 2012). This likely stems from a reduction in the direct contribution of precipitation to streamflow (Li et al., 2009; Mo and Lettenmaier, 2014; Wood and Lettenmaier, 2008), which reduces variability and allows initial conditions to persist for longer. In winter, lower evaporation rates lead to more effective rainfall which 'disrupts' the initial state and limits the skill of ESP forecasts. This is particularly noticeable in flashy catchments with low baseflow contribution, where the hydrological response is driven predominately by rainfall. Under such conditions, rainfall events propagate to streamflow at a much faster rate and memory of initial conditions is lost quickly. At longer lead times, ESP is least skilful when initialised in spring. Both Harrigan et al. (2018) and Svensson (2016) also found lower longer-range skill for forecasts initialised in spring in the UK. The former attributed this to the transition from wet conditions with small soil moisture deficits to dry conditions with large soil moisture deficits. Given that Ireland shares a similar precipitation regime to the UK, and that ESP skill is negatively impacted by high rainfall variability across the forecast period (Harrigan et al., 2018), this is also a plausible explanation for the results observed here.

4.2 Where is ESP skilful?

ESP is most skilful in the Midlands and least skilful in the Border and South-West. The Midlands is a lowland karst region which is underlain by permeable Carboniferous limestone, characterised by several locally and regionally important aquifers. Given that soils in this region are also well-drained, catchments located here have higher storage capacity and hence greater skill due to their long 'memory'. Both the Border and the West are poorly-drained regions, with the former characterised by unproductive bedrock aquifers. This partly explains the low storage capacity of catchments in these regions, which have quick hydrological response times and poor persistence of initial conditions, resulting in lower ESP skill. Similar patterns were noted for persistence forecasts (Foran Quinn et al., 2021).

4.3 Why is ESP skilful?

ESP skill displays a strong relationship with modelled catchment storage capacity and catchment BFI values, with higher skill scores returned for catchments with greater storage. We conclude that storage capacity is primarily responsible for modulating ESP skill. High BFI catchments have flow regimes dominated by slowly released groundwater (Chiverton et al., 2015) and are characterised by longer response times and lower streamflow variability (Sear et al., 1999; Broderick et al., 2016). This is conducive to greater persistence of initial conditions, with water storage in the soil creating a memory effect whereby anomalous conditions can take weeks or months to wane (Ghannam et al., 2014; Harrigan et al., 2018; Li et al., 2009). The

role played by storage capacity is perhaps best illustrated by the fact that ESP skill decays at a much slower rate in catchments with high BFI, especially during summer when streamflow is derived primarily from stored sources. For example, ESP is moderately skilful (CRPSS ≥ 0.25) out a 2-month lead time for the Inny (ID 26021; BFI $= 0.82$) when initialised in July, but shows adequate (non-neutral) performance relative to climatology (CRPSS > 0.05) up to 4 months ahead. Moreover, whilst ESP tends to perform worse outside of summer months, catchments with relatively high SAAR but also high BFI yield above average skill scores in winter, spring, and autumn. In the Slaney (ID 12001; BFI $= 0.67$; SAAR $= 1167\,\mathrm{mm}$), skilful forecasts are possible up to almost a year ahead in January and February, and up to 3–6 months ahead in spring and autumn. This likely stems from the delayed release of precipitation from groundwater stores (van Dijk et al., 2013), which can lead to temporal streamflow dependence for up to a season ahead (Chiverton et al., 2015).

4.4 Potential for operationalising ESP in Ireland

Our benchmarking results establish that ESP, in its traditional formulation, is skilful in a number of different scenarios, sometimes up to several months in advance. We recommend that ESP be used operationally in Ireland, similar to the HOUK (Prudhomme et al., 2017). Skilful streamflow forecasts at short to extended lead times could prove beneficial for water resources management, particularly in areas such as Dublin where water supply systems have been operating close to capacity and face challenges of supply during dry periods. Given that the predictability of summer rainfall is notoriously difficult over northern Europe (Weisheimer and Palmer, 2014), the true utility of ESP may lie in its ability to leverage initial hydrological conditions, particularly in high storage catchments, to skilfully predict streamflow up to a season ahead during dry months. Operationally, skill could be extended further by initialising forecasts more than once a month (e.g., Girons Lopez et al., 2021). As ESP has also been shown to accurately forecast the occurrence of low and high flow events in many catchments up to at least a month in advance, it may also have practical relevance for decision-makers where it can act as an aid in the management of hydrologic extremes.

In the absence of skilful atmospheric forecasts or improved hydrological process representation, historical ESP provides a lower limit of streamflow forecasting skill (Harrigan et al., 2018). However, we show that it is possible to improve ESP skill during winter by conditioning the method on the NAO. Improvements in forecast skill (CRPSS) of 7–18% over lead times of 1 to 3 months are possible in catchments where meteorological conditions are the dominant control on skill. Notwithstanding differences in study design, these improvements are comparable to those of Beckers et al. (2016) using an ENSO-conditioned ESP. We do acknowledge, however, that these improvements are thus limited to specific catchments and are on top of a low initial skill base. In addition to improvements in overall forecast performance, NAO-conditioned ESP increases low flow reliability and extends the lead time over which skilful discrimination of both low and high flow events is possible. As winter is the most important season for groundwater recharge, during which reservoirs fill up to be used over the summer, the ability to more accurately forecast dry winters in this way is extremely valuable for water managers, allowing them to anticipate the water situation beyond what is provided by the forecast alone. Hence, the greatest benefit of NAO-conditioned ESP may be found in its improved low flow reliability and discrimination, rather than its overall performance.

4.5 Potential for future work

ESP skill is to a large extent dependent on the ability of hydrological models to accurately simulate catchment processes (Wang et al., 2011). It follows that further advances in ESP will likely require better representation of initial hydrological conditions and their evolution over time. Model structural and parameter uncertainty are therefore important considerations. Multi-parameter ensembles, data assimilation (e.g., Franz et al., 2014), state updating (e.g., Gibbs et al., 2018), and the use of satellite data and remote sensing are potential ways through which estimates of initial conditions could be improved. It may also be possible to improve predictability by choosing model structures that are more capable of representing key flow pathways (i.e., groundwater, quick flow, etc.) and hence generate more accurate initial states. In this paper, GR4J is used as a parsimonious conceptual model to determine when and where skill is possible. Ongoing work will explore whether additional model complexity adds forecast skill at different initialisation and lead times through the use of models with different structures and parameter dimensionality. In an operational setting, this could be extended to include more spatially discrete physically-based hydrological models that may better account for initial conditions. The additional benefit derived from using ensembles of models for maximising skill persistence could also be assessed for different lead times and initialisation months. This is a promising avenue, as model diversity has been shown to enhance forecast skill in ensemble experiments (Sharma et al., 2019).

We conducted a basic analysis of the relationship between forecast skill and catchment characteristics, using a small selection of descriptors. A more comprehensive investigation of this relationship could be carried out, employing clustering techniques (e.g., Girons Lopez et al., 2021; Pechlivanidis et al., 2020) and a wider range of hydrological signatures. As PCDs are available for a larger sample of 215 catchments, skill could be inferred in areas where modelling is not feasible (e.g., due to sparse or poor-quality observational data) based on *a priori* knowledge of local hydrological conditions. This could also be achieved by regionalising model parameters.

Finally, our use of NAO-conditioned ESP as described in this paper is only one way in which seasonal climate information can be incorporated into ESP forecasts. Whilst we use precipitation analogues derived from GloSea5 hindcasts to generate a new ensemble, an alternative approach is to post-process the historical ESP ensemble, similar to Beckers et al. (2016) or Yuan and Zhu (2018). This would involve sub-selecting ensemble members by comparing the NAO index at the time of forecast with the NAO index on the same day of a year in the historical record (e.g., using correlation analysis or a k-nearest neighbours approach). A different approach could be to condition model parameter sets rather than model inputs. It may also be possible to improve skill outside of winter, as the winter NAO has shown lagged correlations with summer rainfall over Ireland (Murphy et al., 2013) and river flows in the UK (Wilby, 2001). Seasonal forecasts of precipitation and temperature could also be incorporated directly into the process, in so-called climate-model based SHF (Yuan et al., 2015).

5 Conclusions

Ensemble Streamflow Prediction is a popular approach to seasonal hydrological forecasting that remains used some 40 years after first development. Here, we benchmarked ESP skill for a diverse sample of Irish catchments and conclude that it is skilful

against streamflow climatology, but that the level of skill is strongly dependent on lead time, initialisation month, and individual catchment location and storage properties. In summary, we find that:

- ESP skill (CRPSS) decays rapidly as a function of lead time, but the rate of decay is much slower in catchments with high storage capacity where initial conditions alone can provide skill up to several months in advance.
- For short (1–3 days), extended (1–2 weeks), and monthly lead times, ESP is most skilful when initialised during summer and least skilful when initialised during winter. At seasonal and annual lead times, ESP is least skilful when initialised during spring and about as skilful in autumn as it is in summer.
- ESP is most skilful in the Midlands, Mid-West, and East regions of Ireland, where slower responding catchments and the underlying lithology favour high storage capacity and longer hydrological 'memory.'
- ESP is capable of accurately discriminating between events and non-events for both low and high flows up to a month ahead in the majority of catchments. At lead times longer than 1 month, the number of catchments for which discriminaiton is possible depends on initialisation month.
- NAO-conditioned ESP improves winter skill (CRPSS) in fast-responding, low storage catchments in the Border and West regions, where the influence of meteorological forcing outweighs that from initial conditions. These improvements are more substantial over longer lead times of 2 and 3 months when the underlying NAO signal is less obscured by noise.
- NAO-conditioned ESP improves reliability of low flow forecasts in nearly all catchments and reduces reliability of high flow forecasts, except for specific runoff dominated catchments.
- NAO-conditioned ESP extends the lead times over which skilful discrimination of low and high flow events is possible. This is particularly beneficial for forecasting dry winters, which can provide forewarning to water managers about potentially problematic conditions.

We have demonstrated the skill of historical ESP for Ireland and highlighted its utility during the dry season, when demand for outlooks may be greatest. We have also shown how to improve ESP during winter, the season most critical for water managers. In light of the potential benefits for decision-makers, we recommend that ESP and conditioned ESP are operationalised, as they are serious contenders for producing skilful seasonal streamflow forecasts in Ireland.

Data availability. Streamflow data are available from the Office of Public Works (https://waterlevel.ie/) and the Environmental Protection Agency (https://www.epa.ie/hydronet/). Climate data and the ESP hindcast archive are available upon request from the authors. Supplement Table S1 includes metadata for all 46 catchments as well as model parameter values and data used to generate Table 2, Fig. 2, and Fig. 7.

Appendix A: Non-parametric Kling–Gupta efficiency

The non-parametric Kling–Gupta efficiency (KGE_{NP}; Pool et al., 2018) is a modification of the traditional KGE (Gupta et al., 2009) that uses the non-parametric Spearman rank correlation coefficient and normalised flow-duration curves to represent discharge dynamics and discharge variability, respectively. It is defined as:

$$\text{KGE}_{\text{NP}} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{\text{NP}} - 1)^2 + (\rho - 1)^2} \tag{A1}$$

$$\beta = \frac{\mu_{\text{s}}}{\mu_{\text{o}}} \tag{A2}$$

$$\alpha_{\text{NP}} = 1 - \frac{1}{2} \sum_{k=1}^{n} \left| \frac{Q_{\text{s}}(I(k))}{n \times \mu_{\text{s}}} - \frac{Q_{\text{o}}(J(k))}{n \times \mu_{\text{o}}} \right| \tag{A3}$$

Where ρ is the non-parametric Spearman rank correlation coefficient between the simulated and observed time series, μ_{s} and μ_{o} are the mean of the simulated and observed time series, respectively, and $I(k)$ and $J(k)$ are the time steps when the k^{th} largest flow occurs within the simulated and observed time series, respectively. β represents discharge volume. α_{NP} is calculated from the absolute difference between the normalised flow-duration curves.

Appendix B: Continuous ranked probability skill score

The continuous ranked probability score (CRPS; Hersbach, 2000) measures the integrated squared difference between the forecast cumulative distribution function (CDF) and the empirical CDF of the observation. For a continuous random variable X (e.g., streamflow) with probability density function f_X, the CRPS between the forecast CDF, denoted F_X, and the empirical CDF of the observation y, denoted F_y, is defined as:

$$\text{CRPS}(F_X, y) = \int_{-\infty}^{\infty} \left[F_X(x) - F_y(x) \right]^2 dx \tag{B1}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt \tag{B2}$$

$$F_y(x) = H(x - y) \tag{B3}$$

Where H is the Heaviside step function: $H(x) = 1$ for $x \geq 0$ and $H(x) = 0$ for $x < 0$. The continuous ranked probability skill score (CRPSS) is then given by:

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}_{\text{Sys}}}}{\overline{\text{CRPS}_{\text{Ref}}}} \tag{B4}$$

Where $\overline{\text{CRPS}_{\text{Sys}}}$ is the average CRPS of the forecasting system for a set of forecast–observation pairs, and $\overline{\text{CRPS}_{\text{Ref}}}$ is the equivalent for the reference forecast. The CRPSS ranges from $-\infty$ to 1, with positive (negative) values indicating better (worse) performance than the reference forecast.

Author contributions. SD designed the study with input from SH. JK, AAS, and NS contributed the GloSea5 data used to condition the ESP method. SD carried out the modelling, analysed the results, and produced the figures under the supervision of CM. SD prepared the manuscript with contributions from all co-authors.

Competing interests. The authors declare no competing interests.

Acknowledgements. CM, SD, SG, and DFQ gratefully acknowledge funding from Science Foundation Ireland (Grant No. SFI/17/CDA/4783). We thank two anonymous referees for their constructive feedback that has improved this paper.

References

Amnatsan, S., Yoshikawa, S., and Kanae, S.: Improved Forecasting of Extreme Monthly Reservoir Inflow Using an Analogue-Based Forecasting Method: A Case Study of the Sirikit Dam in Thailand, Water, 10, 1614, https://doi.org/10.3390/w10111614, 2018.

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., and Lettenmaier, D. P.: Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments, Water Resour. Res., 52, 4209–4225, https://doi.org/10.1002/2015WR017864, 2016.

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, Hydrol. Earth Syst. Sci., 22, 2057–2072, https://doi.org/10.5194/hess-22-2057-2018, 2018.

Arsenault, R., Brissette, F., and Martel, J.-L.: The hazards of split-sample validation in hydrological model calibration, J. Hydrol., 566, 346–362, https://doi.org/10.1016/j.jhydrol.2018.09.027, 2018.

Baker, D. B., Richards, R. P., Loftus, T. T., and Kramer, J. W.: A NEW FLASHINESS INDEX: CHARACTERISTICS AND APPLICATIONS TO MIDWESTERN RIVERS AND STREAMS, J. Am. Water Resour. Assoc., 40, 503–522, https://doi.org/10.1111/j.1752-1688.2004.tb01046.x, 2004.

Baker, L. H., Shaffrey, L. C., and Scaife, A. A.: Improved seasonal prediction of UK regional precipitation using atmospheric circulation, Int. J. Climatol., 38, e437–e453, https://doi.org/10.1002/joc.5382, 2018.

Bazile, R., Boucher, M.-A., Perreault, L., and Leconte, R.: Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate, Hydrol. Earth Syst. Sci., 21, 5747–5762, https://doi.org/10.5194/hess-21-5747-2017, 2017.

Beckers, J. V. L., Weerts, A. H., Tijdeman, E., and Welles, E.: ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction, Hydrol. Earth Syst. Sci., 20, 3277–3287, https://doi.org/10.5194/hess-20-3277-2016, 2016.

Bell, V. A., Davies, H. N., Kay, A. L., Brookshaw, A., and Scaife, A. A.: A national-scale seasonal hydrological forecast system: development and evaluation over Britain, Hydrol. Earth Syst. Sci., 21, 4681–4691, https://doi.org/10.5194/hess-21-4681-2017, 2017.

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

Berghuijs, W. R., Woods, R. A., and Hrachowitz, M.: A precipitation shift from snow towards rain leads to a decrease in streamflow, Nat. Clim. Chang., 4, 583–586, https://doi.org/10.1038/nclimate2246, 2014.

Bergmeir, C., Molina, D., and Benítez, J. M.: Memetic Algorithms with Local Search Chains in R: The Rmalschains Package, J. Stat. Softw., 75, 1–33, https://doi.org/10.18637/jss.v075.i04, 2016.

Bergmeir, C., Benítez, J. M., Molina, D., Davies, R., Eddelbuettel, D., and Hansen, N.: Rmalschains: Continuous Optimization using Memetic Algorithms with Local Search Chains (MA-LS-Chains) in R, https://CRAN.R-project.org/package=Rmalschains, R package version 0.2-6, 2019.

Bierkens, M. F. P. and van Beek, L. P. H.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, J. Hydrometeorol., 10, 953–968, https://doi.org/10.1175/2009JHM1034.1, 2009.

Bradley, A. A., Habib, M., and Schwartz, S. S.: Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach, Water Resour. Res., 51, 7382–7400, https://doi.org/10.1002/2014WR016811, 2015.

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., and Murphy, C.: Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods, Water Resour. Res., 52, 8343–8373, https://doi.org/10.1002/2016WR018850, 2016.

Broderick, C., Murphy, C., Wilby, R. L., Matthews, T., Prudhomme, C., and Adamson, M.: Using a Scenario-Neutral Framework to Avoid Potential Maladaptation to Future Flood Risk, Water Resour. Res., 55, 1079–1104, https://doi.org/10.1029/2018WR023623, 2019.

Chiverton, A., Hannaford, J., Holman, I., Corstanje, R., Prudhomme, C., Bloomfield, J., and Hess, T. M.: Which catchment characteristics control the temporal dependence structure of daily river flows?, Hydrol. Process., 29, 1353–1369, https://doi.org/10.1002/hyp.10252, 2015.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resour. Res., 48, W05 552, https://doi.org/10.1029/2011WR011721, 2012.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The suite of lumped GR hydrological models in an R package, Environ. Model. Softw., 94, 166–171, https://doi.org/10.1016/j.envsoft.2017.05.002, 2017.

Coron, L., Delaigue, O., Thirel, G., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, https://doi.org/10.15454/EX11NA, R package version 1.4.3.65, 2020.

Day, G. N.: Extended Streamflow Forecasting Using NWSRFS, J. Water Resour. Plan. Manag., 111, 157–170, https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157), 1985.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., and Zhu, Y.: Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmos. Sci. Lett., 11, 114–122, https://doi.org/10.1002/asl.261, 2010.

Demargne, J., Wu, L., Regonda, S. K., Brown, J. D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H. D., Fresch, M., Schaake, J., and Zhu, Y.: The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, Bull. Am. Meteorol. Soc., 95, 79–98, https://doi.org/10.1175/BAMS-D-12-00081.1, 2014.

Dixon, S. G. and Wilby, R. L.: A seasonal forecasting procedure for reservoir inflows in Central Asia, River Res. Appl., 35, 1141–1154, https://doi.org/10.1002/rra.3506, 2019.

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., and Robinson, N.: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world?, Geophys. Res. Lett., 41, 5620–5628, https://doi.org/10.1002/2014GL061146, 2014.

Fan, F. M., Schwanenberg, D., Alvarado, R., Assis dos Reis, A., Collischonn, W., and Naumman, S.: Performance of Deterministic and Probabilistic Hydrological Forecasts for the Short-Term Optimization of a Tropical Hydropower Reservoir, Water Resour. Manag., 30, 3609–3625, https://doi.org/10.1007/s11269-016-1377-8, 2016.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorol. Appl., 15, 19–24, https://doi.org/10.1002/met.45, 2008.

Ficchì, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.-O., Overloop, P.-J. V., and Jay-Allemand, M.: Optimal Operation of the Multireservoir System in the Seine River Basin Using Deterministic and Ensemble Forecasts, J. Water Resour. Plan. Manag., 142, 05015 005, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571, 2016.

Foran Quinn, D., Murphy, C., Wilby, R. L., Matthews, T., Broderick, C., Golian, S., Donegan, S., and Harrigan, S.: Benchmarking seasonal forecasting skill using river flow persistence in Irish catchments, Hydrol. Sci. J., 66, 672–688, https://doi.org/10.1080/02626667.2021.1874612, 2021.

Förster, K., Hanzer, F., Stoll, E., Scaife, A. A., MacLachlan, C., Schöber, J., Huttenlau, M., Achleitner, S., and Strasser, U.: Retrospective forecasts of the upcoming winter season snow accumulation in the Inn headwaters (European Alps), Hydrol. Earth Syst. Sci., 22, 1157–1173, https://doi.org/10.5194/hess-22-1157-2018, 2018.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin, J. Hydrometeorol., 4, 1105–1118, https://doi.org/10.1175/1525-7541(2003)004<1105:VONWSE>2.0.CO;2, 2003.

Franz, K. J., Hogue, T. S., Barik, M., and He, M.: Assessment of SWE data assimilation for ensemble streamflow predictions, J. Hydrol., 519, 2737–2746, https://doi.org/10.1016/j.jhydrol.2014.07.008, 2014.

Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17, 395–407, https://doi.org/10.5194/hess-17-395-2013, 2013.

Ghannam, K., Nakai, T., Paschalis, A., Oishi, C. A., Kotani, A., Igarashi, Y., Kumagai, T., and Katul, G. G.: Persistence and memory timescales in root-zone soil moisture dynamics, Water Resour. Res., 52, 1427–1445, https://doi.org/10.1002/2015WR017983, 2014.

Gibbs, M. S., McInerney, D., Humphrey, G., Thyer, M. A., Maier, H. R., Dandy, G. C., and Kavetski, D.: State updating and calibration period selection to improve dynamic monthly streamflow forecasts for an environmental flow management application, Hydrol. Earth Syst. Sci., 22, 871–887, https://doi.org/10.5194/hess-22-871-2018, 2018.

Girons Lopez, M., Crochemore, L., and Pechlivanidis, I. G.: Benchmarking an operational hydrological model for providing seasonal forecasts in Sweden, Hydrol. Earth Syst. Sci., 25, 1189–1209, https://doi.org/10.5194/hess-25-1189-2021, 2021.

Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic forecasts, calibration and sharpness, J. R. Stat. Soc. B, 69, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x, 2007.

Greuell, W., Franssen, W. H. P., and Hutjes, R. W. A.: Seasonal streamflow forecasts for Europe – Part 2: Sources of skill, Hydrol. Earth Syst. Sci., 23, 371–391, https://doi.org/10.5194/hess-23-371-2019, 2019.

Gupta, H. V., Sorooshian, S., and Yapo, P. O.: Status of Automatic Calibration for Hydrologic Models: Comparison with Multilevel Expert Calibration, J. Hydrol. Eng., 4, 135–143, https://doi.org/10.1061/(ASCE)1084-0699(1999)4:2(135), 1999.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Gustard, A., Bullock, A., and Dixon, J. M.: Low flow estimation in the United Kingdom, Tech. Rep. 108, Institute of Hydrology, Wallingford, UK, http://nora.nerc.ac.uk/id/eprint/6050/1/IH_108.pdf, 1992.

Hamlet, A. F. and Lettenmaier, D. P.: Columbia River Streamflow Forecasting Based on ENSO and PDO Climate Signals, J. Water Resour. Plan. Manag., 125, 333–341, https://doi.org/10.1061/(ASCE)0733-9496(1999)125:6(333), 1999.

Hamlet, A. F., Huppert, D., and Lettenmaier, D. P.: Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower, J. Water Resour. Plan. Manag., 128, 91–101, https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(91), 2002.

Hansen, N. and Ostermeier, A.: Completely Derandomized Self-Adaptation in Evolution Strategies, Evol. Comput., 9, 159–195, https://doi.org/10.1162/106365601750190398, 2001.

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., and Tanguy, M.: Benchmarking ensemble streamflow prediction skill in the UK, Hydrol. Earth Syst. Sci., 22, 2023–2039, https://doi.org/10.5194/hess-22-2023-2018, 2018.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrol. Sci. J., 31, 13–24, https://doi.org/10.1080/02626668609491024, 1986.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, https://doi.org/10.5194/hess-11-1267-2007, 2007.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res. Atmos., 114, D04 114, https://doi.org/10.1029/2008JD010969, 2009.

Luo, L. and Wood, E. F.: Monitoring and predicting the 2007 U.S. drought, Geophys. Res. Lett., 34, L22 702, https://doi.org/10.1029/2007GL031673, 2007.

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.: Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system, Q. J. R. Meteorol. Soc., 141, 1072–1084, https://doi.org/10.1002/qj.2396, 2015.

Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, Weather Forecast., 14, 713–725, https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2, 1999.

Meißner, D., Klein, B., and Ionita, M.: Development of a monthly to seasonal forecast framework tailored to inland waterway transport in central Europe, Hydrol. Earth Syst. Sci., 21, 6401–6423, https://doi.org/10.5194/hess-21-6401-2017, 2017.

MeteoSwiss: easyVerification: Ensemble Forecast Verification for Large Data Sets, https://CRAN.R-project.org/package=easyVerification, R package version: 0.4.4, 2017.

Mills, P., Nicholson, O., and Reed, D.: Flood Studies Update Technical Research Report: Volume IV – Physical Catchment Descriptors, Tech. rep., Office of Public Works, Trim, Ireland, https://opw.hydronet.com/data/files/Technical%20Research%20Report%20-%20Volume%20IV%20-%20Physical%20Catchment%20Descriptors.pdf, 2014.

Mo, K. C. and Lettenmaier, D. P.: Hydrologic Prediction over the Conterminous United States Using the National Multi-Model Ensemble, J. Hydrometeorol., 15, 1457–1472, https://doi.org/10.1175/JHM-D-13-0197.1, 2014.

Molina, D., Lozano, M., García-Martínez, C., and Herrera, F.: Memetic Algorithms for Continuous Optimisation Based on Local Search Chains, Evol. Comput., 18, 27–63, https://doi.org/10.1162/evco.2010.18.1.18102, 2010.

Murphy, C., Harrigan, S., Hall, J., and Wilby, R. L.: Climate-driven trends in mean and high flows from a network of reference stations in Ireland, Hydrol. Sci. J., 58, 755–772, https://doi.org/10.1080/02626667.2013.782407, 2013.

Mushtaq, S., Chen, C., Hafeez, M., Maroulis, J., and Gabriel, H.: The economic value of improved agrometeorological information to irrigators amid climate variability, Int. J. Climatol., 32, 567–581, https://doi.org/10.1002/joc.2015, 2012.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Neumann, J. L., Arnal, L., Emerton, R. E., Griffith, H., Hyslop, S., Theofanidi, S., and Cloke, H. L.: Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity, Geosci. Commun., 1, 35–57, https://doi.org/10.5194/gc-1-35-2018, 2018.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, J. Hydrol., 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Pagano, T., Hapuarachchi, P., and Wang, Q.: Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation, Tech. Rep. EP103545, CSIRO: Water for a Healthy Country National Research Flagship, Canberra, Australia, https://doi.org/10.4225/08/58542c672dd2c, 2010.

Paiva, R. C. D., Collischonn, W., Bonnet, M. P., and de Gonçalves, L. G. G.: On the sources of hydrological prediction uncertainty in the Amazon, Hydrol. Earth Syst. Sci., 16, 3127–3137, https://doi.org/10.5194/hess-16-3127-2012, 2012.

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., and Thielen, J.: The monetary benefit of early flood warnings in Europe, Environ. Sci. Policy, 51, 278–291, https://doi.org/10.1016/j.envsci.2015.04.016, 2015a.

Pappenberger, F., Ramos, M., Cloke, H., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., and Salamon, P.: How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, J. Hydrol., 522, 697–713, https://doi.org/10.1016/j.jhydrol.2015.01.024, 2015b.

Pechlivanidis, I. G., Crochemore, L., Rosberg, J., and Bosshard, T.: What Are the Key Drivers Controlling the Quality of Seasonal Streamflow Forecasts?, Water Resour. Res., 56, e2019WR026 987, https://doi.org/10.1029/2019WR026987, 2020.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrol. Sci. J., 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.

Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H., Davis, R., Mackay, J., McKenzie, A., Rudd, A., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A.: Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales, Hydrol. Sci. J., 62, 2753–2768, https://doi.org/10.1080/02626667.2017.1395032, 2017.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05 521, https://doi.org/10.1029/2009WR008328, 2010.

Robertson, D., Bennett, J., and Schepen, A.: How good is my forecasting method? Some thoughts on forecast evaluation using cross-validation based on Australian experiences, https://hepex.inrae.fr/how-good-is-my-forecasting-method-some-thoughts-on-forecast-evaluation-using-cross-validation-based-on-australian-experiences/, (last access: 15 November 2020), 2016.

Santos, L., Thirel, G., and Perrin, C.: Continuous state-space representation of a bucket-type rainfall-runoff model: a case study with the GR4 model using state-space GR4 (version 1.0), Geosci. Model Dev., 11, 1591–1605, https://doi.org/10.5194/gmd-11-1591-2018, 2018.

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

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.: Skillful long-range prediction of European and North American winters, Geophys. Res. Lett., 41, 2514–2519, https://doi.org/10.1002/2014GL059637, 2014.

Schepen, A. and Wang, Q. J.: Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia, Water Resour. Res., 51, 1797–1812, https://doi.org/10.1002/2014WR016163, 2015.

Sear, D. A., Armitage, P. D., and Dawson, F. H.: Groundwater dominated rivers, Hydrol. Process., 13, 255–276, https://doi.org/10.1002/(SICI)1099-1085(19990228)13:3<255::AID-HYP737>3.0.CO;2-Y, 1999.

Sharma, S., Siddique, R., Reed, S., Ahnert, P., and Mejia, A.: Hydrological Model Diversity Enhances Streamflow Forecast Skill at Short- to Medium-Range Timescales, Water Resour. Res., 55, 1510–1530, https://doi.org/10.1029/2018WR023197, 2019.

Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrol. Earth Syst. Sci., 17, 2781–2796, https://doi.org/10.5194/hess-17-2781-2013, 2013.

Singh, S. K.: Long-term Streamflow Forecasting Based on Ensemble Streamflow Prediction Technique: A Case Study in New Zealand, Water Resour. Manag., 30, 2295–2309, https://doi.org/10.1007/s11269-016-1289-7, 2016.

Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L. F., Caron, L.-P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., Kimoto, M., Merryfield, W. J., Mignot, J., Mochizuki, T., Modali, K., Monerie, P.-A., Müller, W. A., Nicolí, D., Ortega, P., Pankatz, K., Pohlmann, H., Robson, J., Ruggieri, P., Sospedra-Alfonso, R., Swingedouw, D., Wang, Y., Wild, S., Yeager, S., Yang, X., and Zhang, L.: North Atlantic climate far more predictable than models imply, Nature, 583, 796–800, https://doi.org/10.1038/s41586-020-2525-0, 2020.

Staudinger, M. and Seibert, J.: Predictability of low flow – An assessment with simulation experiments, J. Hydrol., 519, 1383–1393, https://doi.org/10.1016/j.jhydrol.2014.08.061, 2014.

Steirou, E., Gerlitz, L., Apel, H., and Merz, B.: Links between large-scale circulation patterns and streamflow in Central Europe: A review, J. Hydrol., 549, 484–500, https://doi.org/10.1016/j.jhydrol.2017.04.003, 2017.

Stringer, N., Knight, J., and Thornton, H.: Improving Meteorological Seasonal Forecasts for Hydrological Modeling in European Winter, J. Appl. Meteorol. Climatol., 59, 317–332, https://doi.org/10.1175/JAMC-D-19-0094.1, 2020.

Svensson, C.: Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues, Hydrol. Sci. J., 61, 19–35, https://doi.org/10.1080/02626667.2014.992788, 2016.

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.: Long-range forecasts of UK winter hydrology, Environ. Res. Lett., 10, 064 006, https://doi.org/10.1088/1748-9326/10/6/064006, 2015.

Tang, Q., Zhang, X., Duan, Q., Huang, S., Yuan, X., Cui, H., Li, Z., and Liu, X.: Hydrological monitoring and seasonal forecasting: Progress and perspectives, J. Geogr. Sci., 26, 904–920, https://doi.org/10.1007/s11442-016-1306-z, 2016.

Thielen, J., Bartholmes, J., Ramos, M.-H., and de Roo, A.: The European Flood Alert System – Part 1: Concept and development, Hydrol. Earth Syst. Sci., 13, 125–140, https://doi.org/10.5194/hess-13-125-2009, 2009.

Turner, S. W. D., Bennett, J. C., Robertson, D. E., and Galelli, S.: Complex relationship between seasonal streamflow forecast skill and value in reservoir operations, Hydrol. Earth Syst. Sci., 21, 4841–4859, https://doi.org/10.5194/hess-21-4841-2017, 2017.

Twedt, T. M., Schaake Jr., J. C., and L., P. E.: National weather service extended streamflow prediction, in: Proceedings of the 45th Annual Western Snow Conference, pp. 52–57, Western Snow Conference, Albuquerque, New Mexico, https://westernsnowconference.org/sites/westernsnowconference.org/PDFs/1977Twedt.pdf, 1977.

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

Vaze, J., Chiew, F. H. S., Perraud, J. M., Viney, N., Post, D., Teng, J., Wang, B., Lerat, J., and Goswami, M.: Rainfall-Runoff Modelling Across Southeast Australia: Datasets, Models and Results, Australas. J. Water Resour., 14, 101–116, https://doi.org/10.1080/13241583.2011.11465379, 2011.

Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P.: How seasonal forecast could help a decision maker: an example of climate service for water resource management, Adv. Sci. Res., 13, 51–55, https://doi.org/10.5194/asr-13-51-2016, 2016.

Walsh, S.: New long-term rainfall averages for Ireland, in: Irish National Hydrology Conference 2012, pp. 3–12, Hydrology Ireland, Tullamore, Ireland, http://hydrologyireland.ie/wp-content/uploads/2016/11/01-Walsh-New-Long-Term-Rainfall-Averages-for-Ireland-1.pdf, 2012.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., and Wood, E. F.: Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System, J. Hydrometeorol., 20, 99–115, https://doi.org/10.1175/JHM-D-18-0040.1, 2019.

Wang, E., Zhang, Y., Luo, J., Chiew, F. H. S., and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05 516, https://doi.org/10.1029/2010WR009922, 2011.

Watts, G., von Christierson, B., Hannaford, J., and Lonsdale, K.: Testing the resilience of water supply systems to long droughts, J. Hydrol., 414–415, 255–267, https://doi.org/10.1016/j.jhydrol.2011.10.038, 2012.

Wedgbrow, C. S., Wilby, R. L., Fox, H. R., and O'Hare, G.: Prospects for seasonal forecasting of summer drought and low river flow anomalies in England and Wales, Int. J. Climatol., 22, 219–236, https://doi.org/10.1002/joc.735, 2002.

Weisheimer, A. and Palmer, T. N.: On the reliability of seasonal climate forecasts, J. R. Soc. Interface, 11, 20131 162, https://doi.org/10.1098/rsif.2013.1162, 2014.

Werner, K., Brandon, D., Clark, M., and Gangopadhyay, S.: Climate Index Weighting Schemes for NWS ESP-Based Seasonal Volume Forecasts, J. Hydrometeorol., 5, 1076–1090, https://doi.org/10.1175/JHM-381.1, 2004.

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

Wilby, R. L.: Seasonal Forecasting of River Flows in the British Isles Using North Atlantic Pressure Patterns, Water Environ. J., 15, 56–63, https://doi.org/10.1111/j.1747-6593.2001.tb00305.x, 2001.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 4th edition, Elsevier, Amsterdam, 2019.

Wood, A. W. and Lettenmaier, D. P.: A Test Bed for New Seasonal Hydrologic Forecasting Approaches in the Western United States, Bull. Am. Meteorol. Soc., 87, 1699–1712, https://doi.org/10.1175/BAMS-87-12-1699, 2006.

Wood, A. W. and Lettenmaier, D. P.: An ensemble approach for attribution of hydrologic prediction uncertainty, Geophys. Res. Lett., 35, L14 401, https://doi.org/10.1029/2008GL034648, 2008.

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., and Clark, M.: Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, J. Hydrometeorol., 17, 651–668, https://doi.org/10.1175/JHM-D-14-0213.1, 2016.

Yang, L., Tian, F., Sun, Y., Yuan, X., and Hu, H.: Attribution of hydrologic forecast uncertainty within scalable forecast windows, Hydrol. Earth Syst. Sci., 18, 775–786, https://doi.org/10.5194/hess-18-775-2014, 2014.

Yuan, X. and Zhu, E.: A First Look at Decadal Hydrological Predictability by Land Surface Ensemble Simulations, Geophys. Res. Lett., 45, 2362–2369, https://doi.org/10.1002/2018GL077211, 2018.

Yuan, X., Wood, E. F., and Ma, Z.: A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development, WIREs Water, 2, 523–536, https://doi.org/10.1002/wat2.1088, 2015.

Yuan, X., Ma, F., Wang, L., Zheng, Z., Ma, Z., Ye, A., and Peng, S.: An experimental seasonal hydrological forecasting system over the Yellow River basin – Part 1: Understanding the role of initial hydrological conditions, Hydrol. Earth Syst. Sci., 20, 2437–2451, https://doi.org/10.5194/hess-20-2437-2016, 2016.

Zhao, T. and Zhao, J.: Joint and respective effects of long- and short-term forecast uncertainties on reservoir operations, J. Hydrol., 517, 83–94, https://doi.org/10.1016/j.jhydrol.2014.04.063, 2014.

Table 1. Physical catchment descriptors referred to in this study.

Descriptor	Explanation	Units	Range
BFI	Baseflow index; proportion of runoff derived from stored sources	–	0–1
RBI	Richards–Baker flashiness index; oscillations in flow relative to total flow	–	0–1
RR	Runoff ratio; ratio of runoff to received precipitation	–	0–1
AREA	Catchment area	km^2	–
SAAR	Standard-period (1961–1990) average annual rainfall	mm	–
FLATWET	Proportion of time soils expected to be typically quite wet	–	0–1
PEAT	Proportional extent of catchment area classified as peat bog	–	0–1
FOREST	Proportional extent of forest cover	–	0–1
MSL	Main-stream length	km	–
S1085	Slope of main stream excluding the bottom 10% and top 15% of its length	$\mathrm{m\,km}^{-1}$	–
TAYSLO	Taylor–Schwartz measure of mainstream slope	$\mathrm{m\,km}^{-1}$	–

Table 2. Summary statistics of eight catchment characteristics for Ireland and each NUTS III region. The median across n catchments is given with the 5th and 95th percentile ranges in parentheses. Mean annual runoff ($\overline{Q}$), precipitation ($\overline{P}$), and potential evaporation ($\overline{PE}$) were calculated over calendar years 1992–2017. $\overline{F_s}$ [a] is the long-term (calendar years 1992–2017) mean fraction of precipitation falling as snow.

Region	n	Area (km^2)	$\overline{Q}$ ($mm\,yr^{-1}$)	$\overline{P}$ ($mm\,yr^{-1}$)	$\overline{PE}$ ($mm\,yr^{-1}$)	BFI (–)	RBI (–)	RR (–)	$\overline{F_s}$ (–)
IE	46	412	686	1149	565	0.59	0.19	0.62	0.02
		(23, 2286)	(431, 1336)	(905, 1861)	(529, 580)	(0.34, 0.75)	(0.07, 0.5)	(0.5, 0.82)	(0.01, 0.02)
B	6	180	970	1484	540	0.43	0.24	0.73	0.02
		(94, 1279)	(569, 1371)	(1088, 1878)	(521, 551)	(0.3, 0.72)	(0.07, 0.55)	(0.59, 0.83)	(0.02, 0.02)
E	8	290	483	926	560	0.62	0.15	0.55	0.02
		(7, 2193)	(385, 750)	(891, 1149)	(535, 574)	(0.44, 0.72)	(0.12, 0.45)	(0.47, 0.7)	(0.01, 0.02)
MW	10	606	697	1177	571	0.58	0.2	0.64	0.02
		(225, 1891)	(506, 900)	(1043, 1373)	(561, 585)	(0.45, 0.67)	(0.09, 0.36)	(0.5, 0.75)	(0.01, 0.02)
M	6	360	524	986	561	0.71	0.13	0.56	0.02
		(38, 1147)	(440, 644)	(914, 1125)	(556, 566)	(0.53, 0.8)	(0.08, 0.26)	(0.52, 0.62)	(0.02, 0.02)
SE	6	738	644	1085	567	0.56	0.26	0.58	0.02
		(145, 2397)	(473, 1044)	(981, 1325)	(545, 576)	(0.42, 0.66)	(0.19, 0.45)	(0.51, 0.85)	(0.02, 0.03)
SW	6	603	929	1581	569	0.44	0.4	0.71	0.02
		(269, 1206)	(668, 1500)	(1417, 1987)	(567, 574)	(0.34, 0.61)	(0.13, 0.5)	(0.65, 0.8)	(0.02, 0.02)
W	4	308	1046	1512	552	0.6	0.18	0.7	0.01
		(87, 1749)	(723, 1223)	(1198, 1695)	(545, 563)	(0.32, 0.75)	(0.1, 0.54)	(0.63, 0.76)	(0.01, 0.01)

[a] $\overline{F_s}$ was calculated following Berghuijs et al. (2014) where precipitation on days with an average temperature greater than or equal to 1 °C was considered entirely rainfall and precipitation on days with an average temperature below 1 °C was considered entirely snowfall.

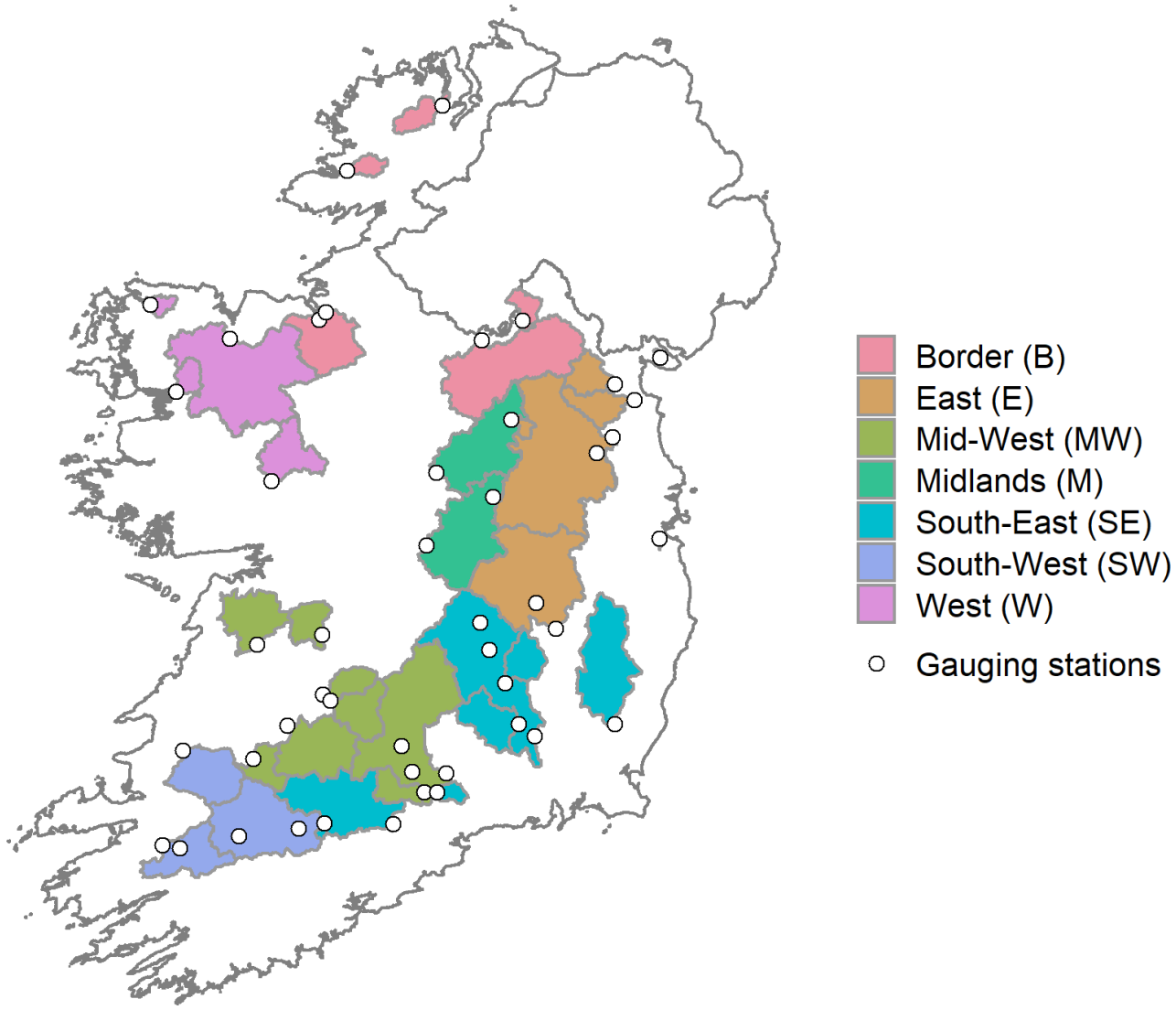

Figure 1. Location of the 46 study catchments, shaded by region, and associated gauging stations (white dots).

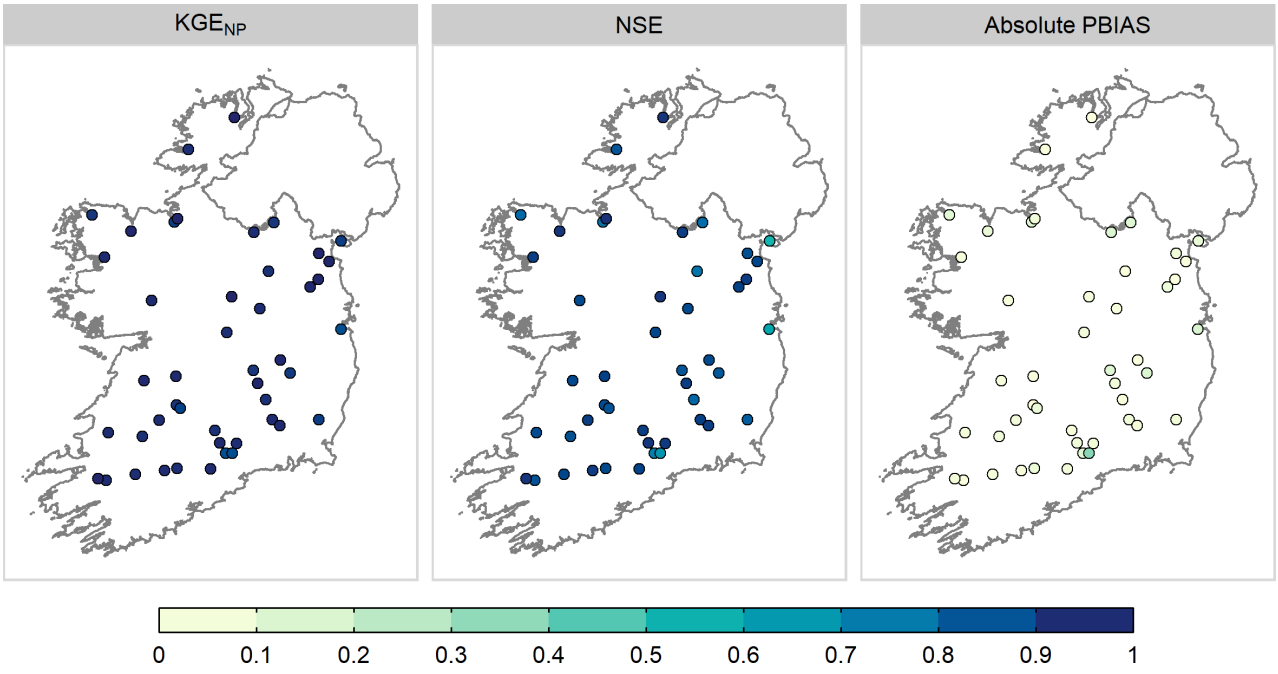

Figure 2. GR4J model performance over the complete period (1993–2017) as measured by KGE_{NP} (left), NSE (middle), and absolute PBIAS (right).

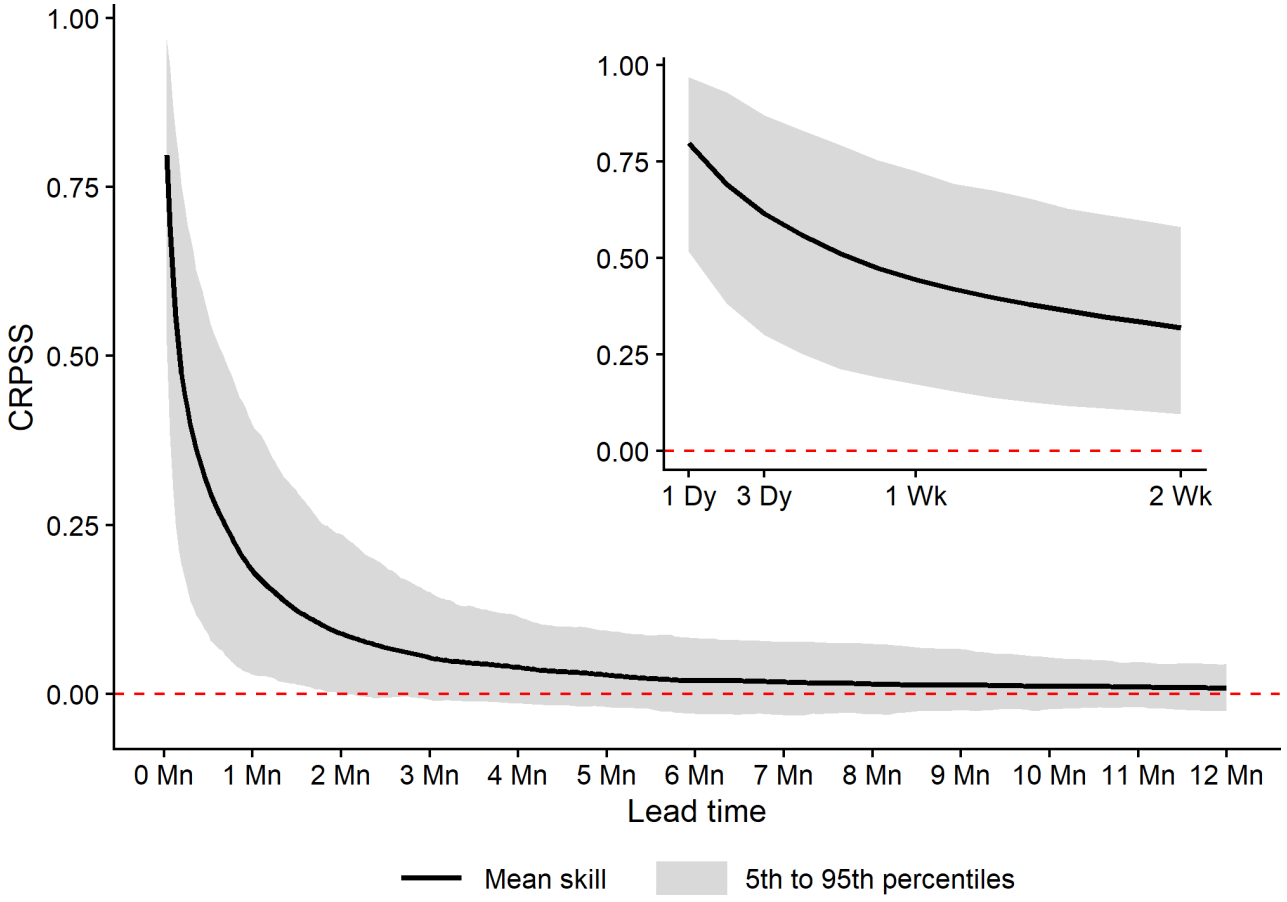

Figure 3. Mean ESP CRPSS values across all 46 study catchments, 12 forecast initialisation months, and all 365 lead times, with short and extended lead times shown inset for readability. Variations in skill scores across all catchments at each lead time are given by the 5th and 95th percentile ensemble range.

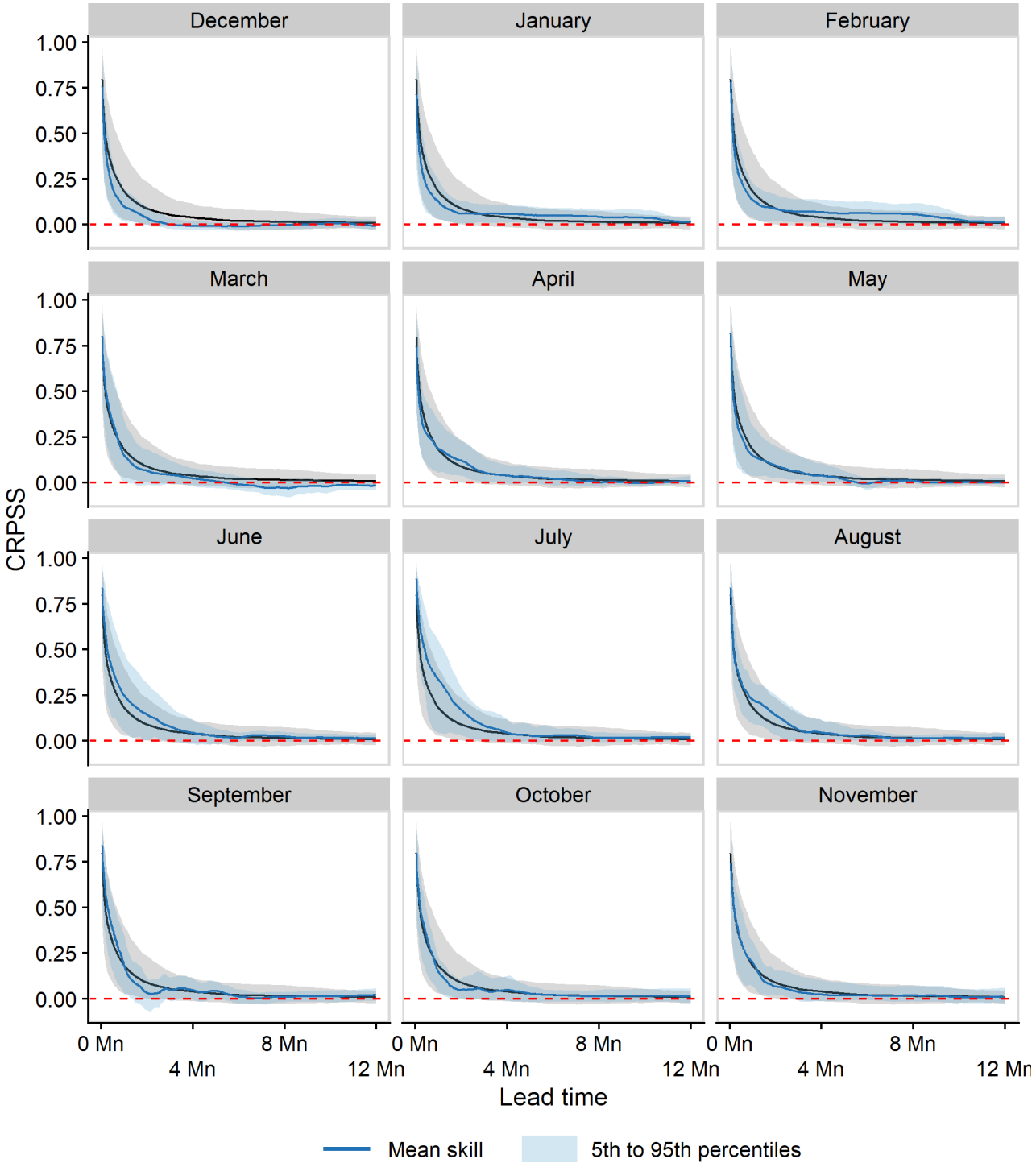

Figure 4. As in Fig. 3, but for each forecast initialisation month. Data from Fig. 3 is included in the background of each panel for reference.

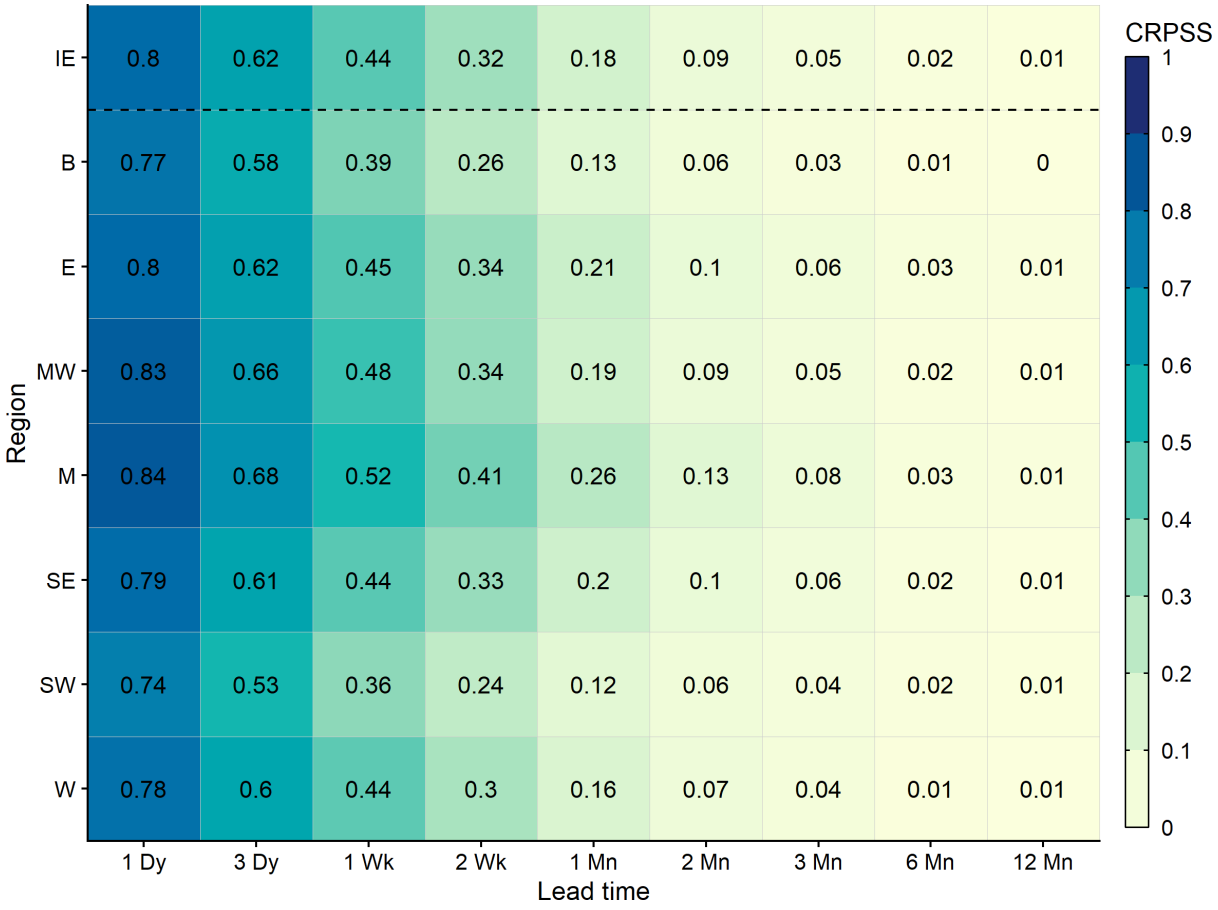

Figure 5. CRPSS values for Ireland (IE) and seven NUTS III regions (B, E, MW, M, SE, SW, and W) averaged across all initialisation months for a selection of lead times: short (1- and 3-day), extended (1- and 2-week), monthly (1- and 2-month), seasonal (3- and 6-month) and annual (12-month).

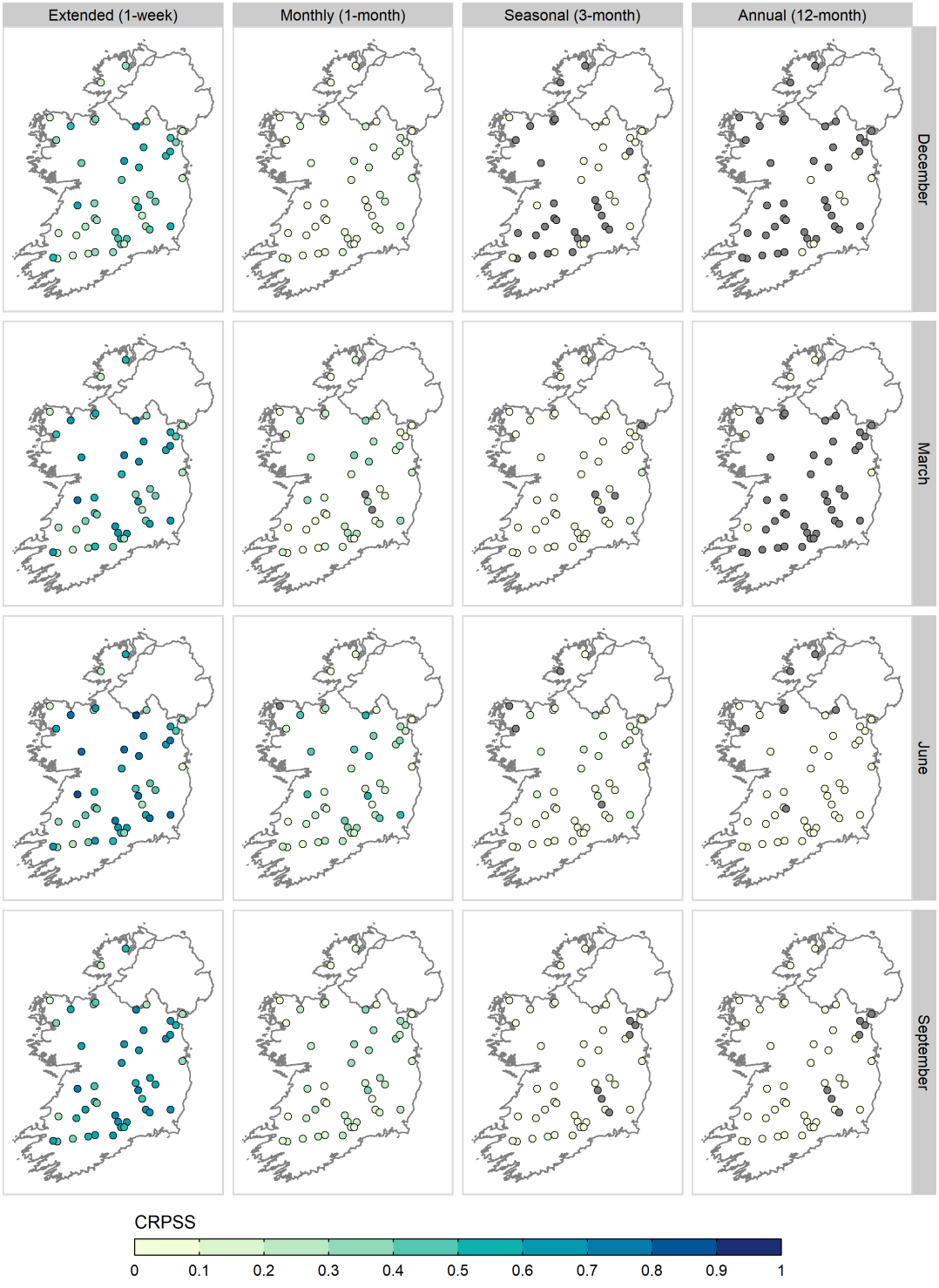

Figure 6. ESP skill for individual forecasts made at the 46 catchments for four sample lead times (columns) and four initialisation months (rows). Catchments with negative skill (CRPSS < 0) are greyed out.

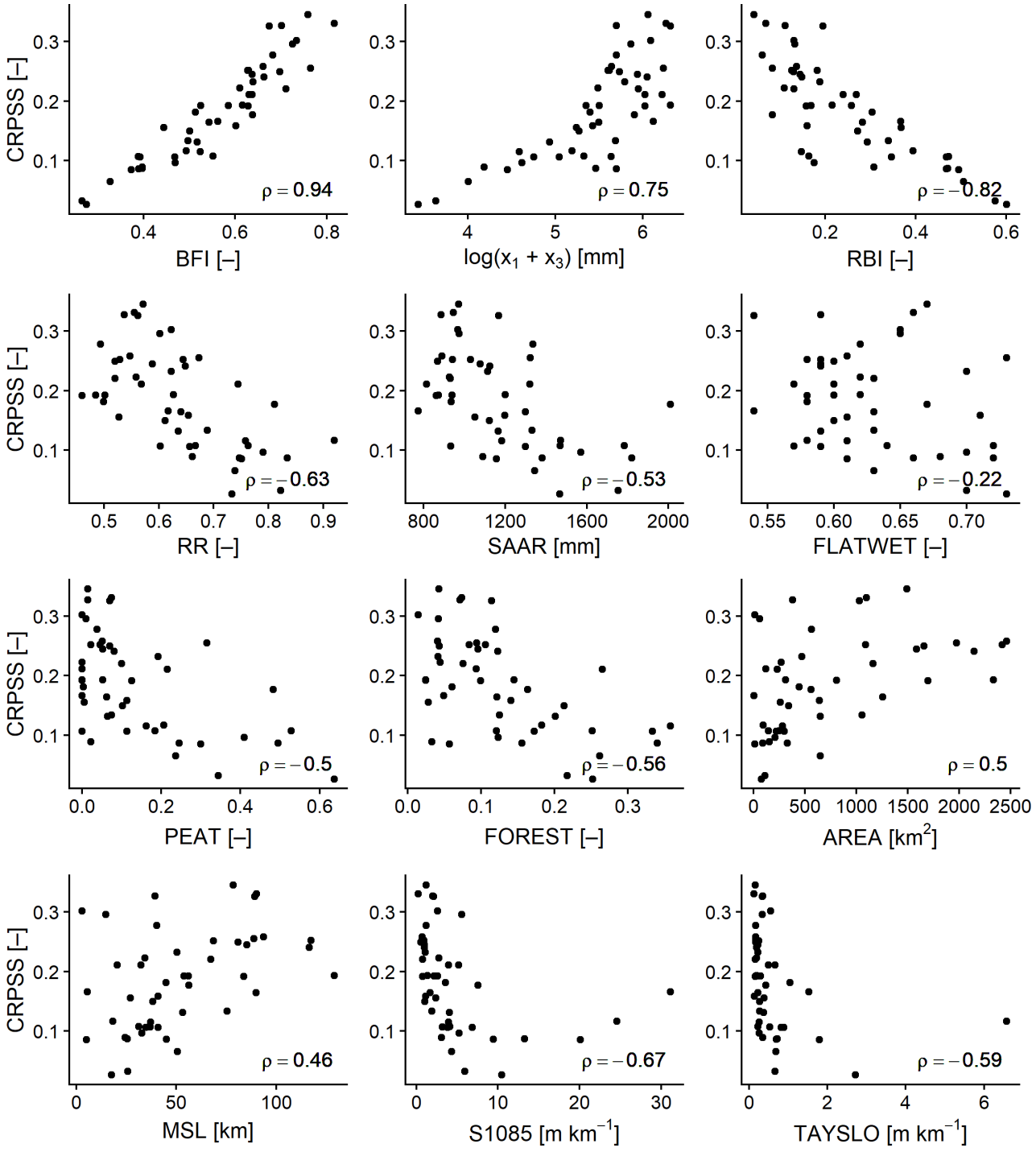

Figure 7. Relationship between 1-month ESP skill (CRPSS) and selected catchment descriptors.

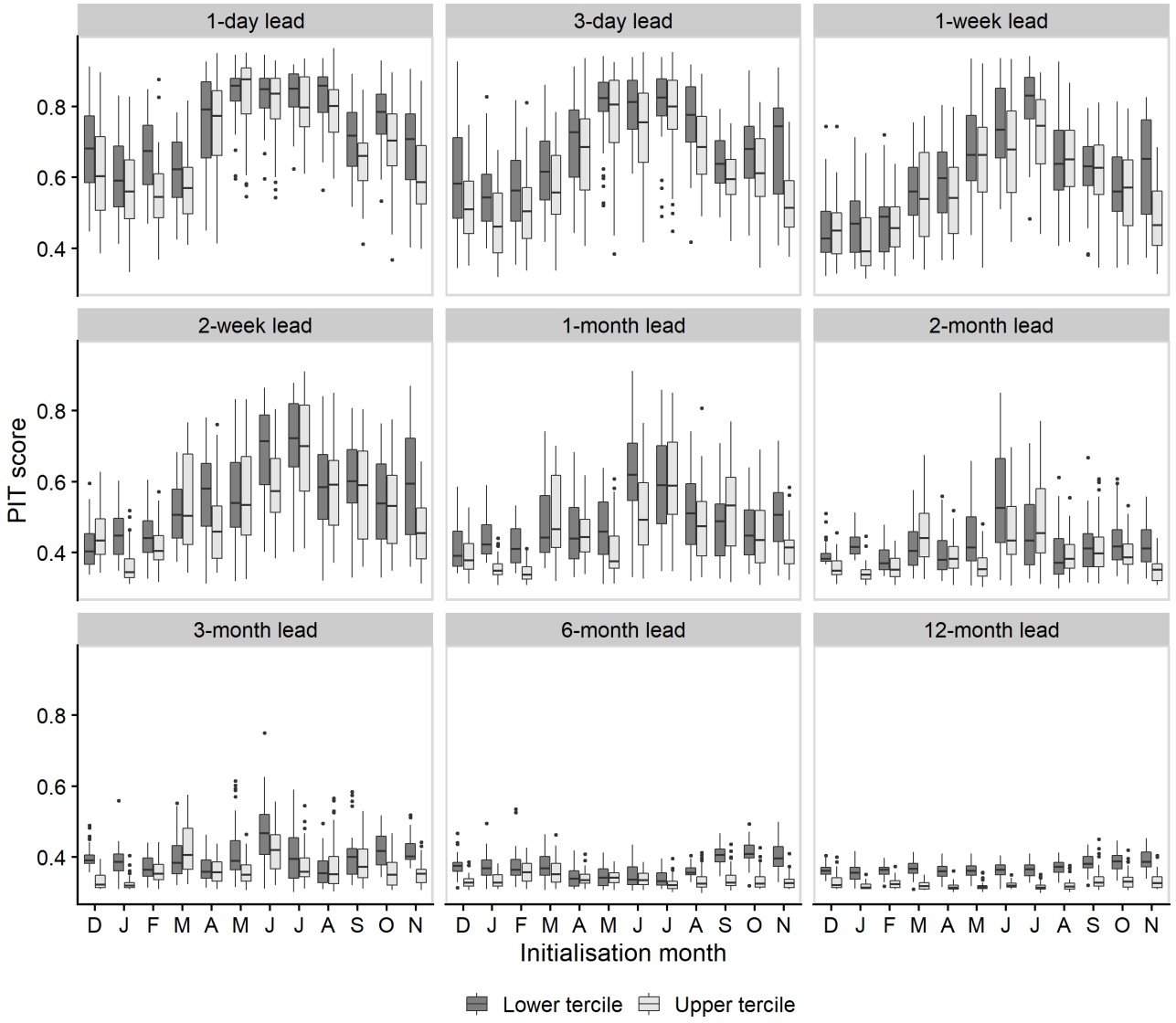

Figure 8. Distribution of PIT score values across all 46 study catchments for each initialisation month and the same selection of lead times as in Fig. 5.

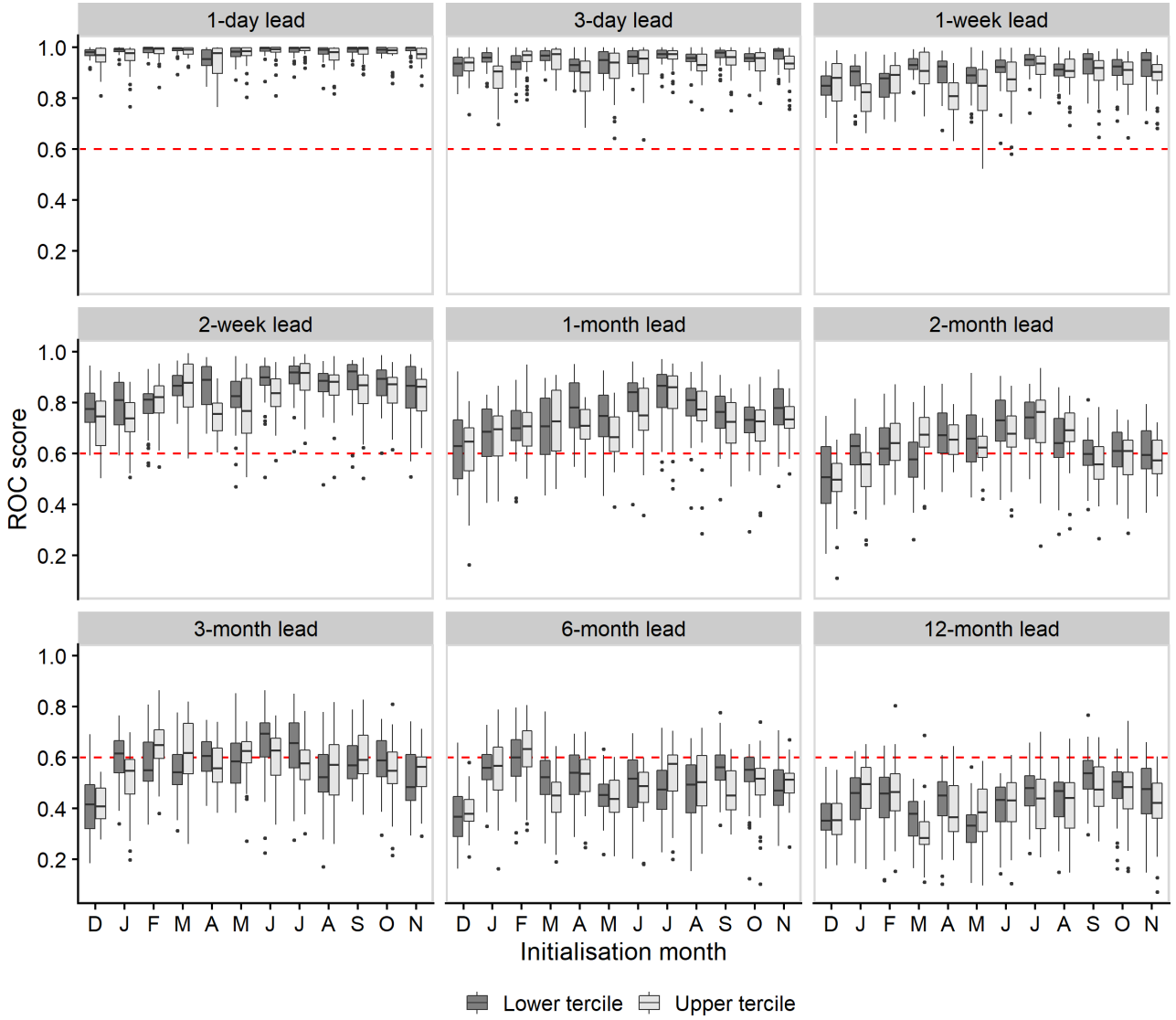

Figure 9. As in Fig. 8, but for the ROC score. The red line denotes the stricter skill threshold of 0.6.

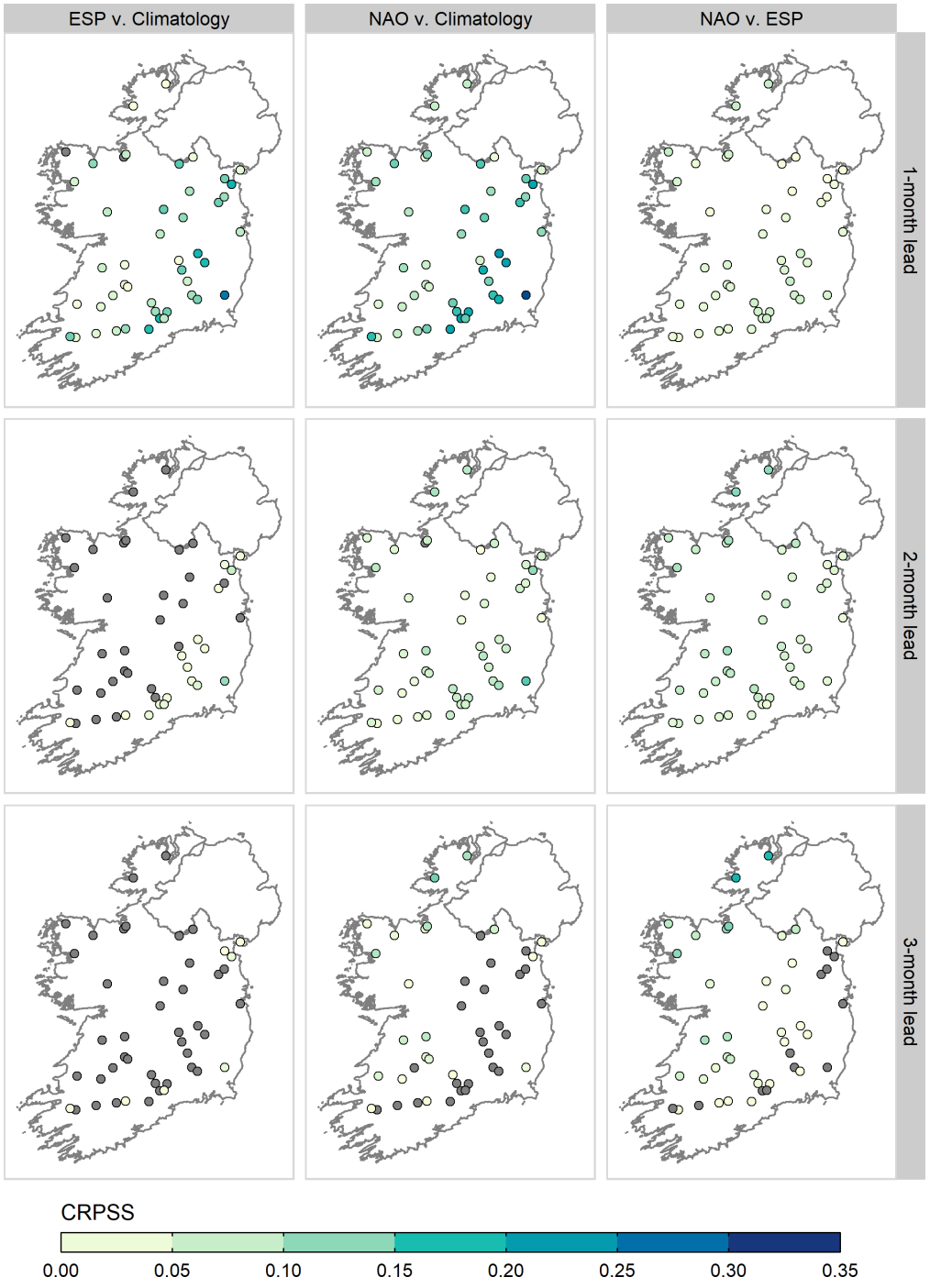

Figure 10. CRPSS values for historical ESP (left column), NAO-conditioned ESP (middle column), and the improvement made by NAO-conditioned ESP over historical ESP (right column), at lead times of 1, 2, and 3 months (rows). Catchments with negative skill (CRPSS $<$ 0) are greyed out.

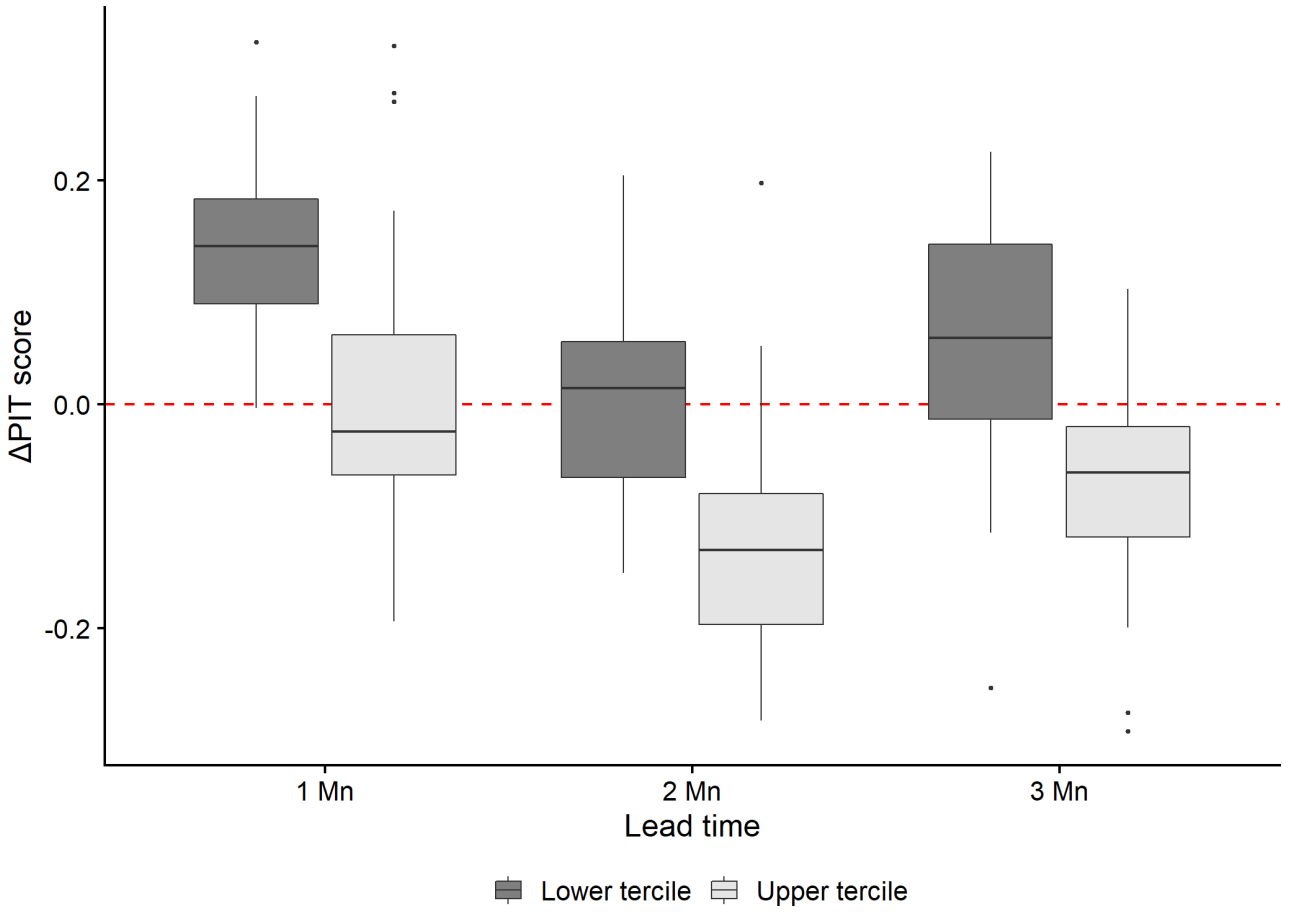

Figure 11. Difference in PIT score values between NAO-conditioned ESP and historical ESP at lead times of 1, 2, and 3 months. Negative values indicate a reduction in reliability whereas positive values indicate an increase in reliability over historical ESP.

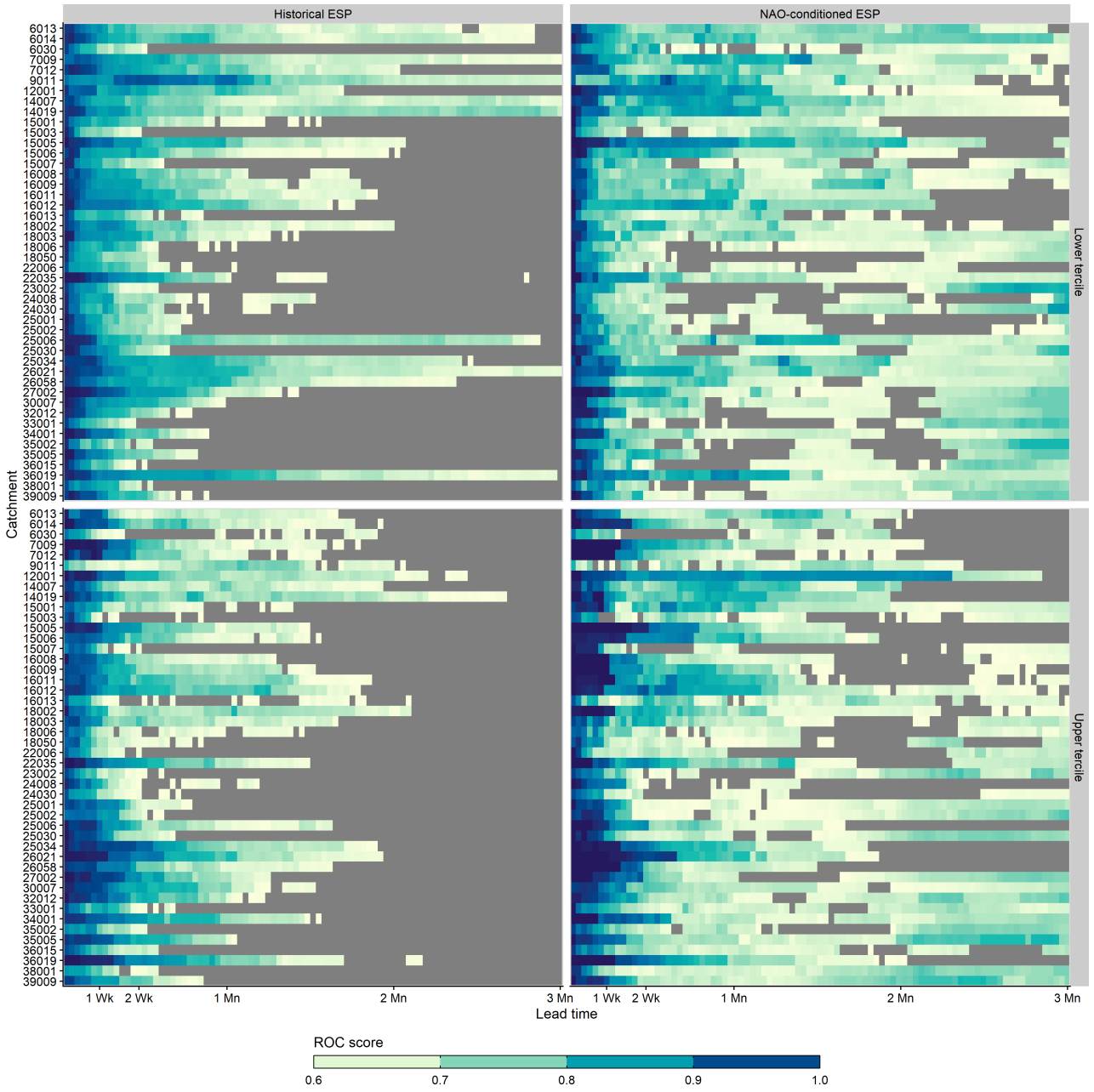

Figure 12. Comparison of ROC scores achieved by historical ESP (left column) and NAO-conditioned ESP (right column) across all 46 study catchments and all lead times for low flow (lower tercile, upper row) and high flow (upper tercile, lower row) events. Cells with no skill (ROC < 0.6) are greyed out.