03 - Seasonal Hydrological Forecasting: Appraising the skill of multiple methods in Ireland

Conor Murphy, Seán Donegan, Dáire Foran Quinn, Saeed Golian

Irish Climate Analysis and Research UnitS (ICARUS), Department of Geography, Maynooth University, Ireland.

ABSTRACT

There is growing evidence that river flows are predictable at monthly to seasonal timescales with operational Seasonal Hydrological Forecasting (SHF) now being trialled and/or operational in some European Countries (e.g., UK Hydrological Outlook at CEH Wallingford). The growing importance and increasing feasibility of SHF globally is evidenced by the work of HEPEX and interest in large-scale SHF as part of climate services (e.g., Copernicus Climate Change Programme (C3S)). Understanding how best to apply the range of tools available and understanding of hydro-climatic processes to develop skilful forecasts requires scientific approaches at the catchment scale. Recent international assessments have shown that there is no single accepted 'best' method for SHF despite several proving useful in different contexts. SHF skill is determined by two major factors – initial hydrological conditions (IHCs) within the catchment and future climate or boundary forcings (for monthly to seasonal forecasts). A range of statistical and model-based approaches for capturing IHCs and boundary conditions have been developed. Research being undertaken by the Science Foundation Ireland funded project HydroCast is evaluating the skill of multiple approaches to SHF in Ireland. This paper summarises the insights and results generated from this work and published in the international literature to date, focusing on the potential skill that can be leveraged from simple to complex approaches including i) empirical/statistical, ii) Ensemble streamflow prediction, and; iii) dynamically informed methods for producing seasonal hydrological forecasts for Ireland. This paper demonstrates when and where different approaches to SHF are skilful and opportunities for operationalising such tools in Ireland.

1. INTRODUCTION

Seasonal Hydrological Forecasting (SHF) systems that leverage understanding of hydro-climatic processes are being used operationally in many countries to predict long-term riverflows to ensure resilience within the water sector. SHF skill is determined by two major factors – initial hydrological conditions (IHCs) within the catchment and future climate or boundary forcings (for monthly to seasonal forecasts) (Wood and Lettenmaier, 2006). Operational SHF systems already exist in some countries. For example, the Hydrological Outlook has been delivering streamflow and groundwater forecasts for UK regions on a monthly basis since June 2013 (Prudhomme *et al*. 2017). Following the summer drought of 2018, water managers in Ireland expressed interest in the potential application of SHF. However, to the authors' knowledge, no previous research has assessed the potential of SHF, or the value of different forecast methods for Ireland. The SFI funded project, HydroCast, aims to establish the skill of SHF methods of varying complexity for Irish conditions. With Ireland's unique hydrology and varied climate, the project provides a test bed for ground-truthing cutting-edge approaches and developing novel methods for merging skill from the most promising models. Specifically, the research sought to demonstrate the predictability of river flows and precipitation at various lead times across seasons, isolate the sources of predictability, the additional skill that can be derived in moving from simple to complex methods and the potential of hybrid methods to extract maximum predictability at various lead times (from weeks to months). This paper provides an overview of the key findings from the HydroCast project to date. We first provide insights into where and when simple approaches such as persistence-based forecasting can add value for water management. Next, we examine the potential of ensemble streamflow prediction (ESP) to capture predictive skill from initial hydrological conditions and the added value of conditioning ESP outputs using dynamical model forecasts of the winter North Atlantic Oscillation (NAO). Finally, we present advances in forecasting seasonal precipitation using hybrid statistical-dynamical methods that leverage skill of dynamical forecasting systems in capturing sea level pressure. While only an overview of each method can be provided here, we encourage the interested reader to explore the open access published papers that provide full details on each method presented (Foran Quinn et al., 2021; Donegan et al., 2021; Golian et al., 2022, respectively).

2. PERSISTENCE BASED METHODS FOR FLOW FORECASTING

When establishing seasonal forecasting capability, it makes sense to begin with simple approaches. SHF based on river flow persistence uses the most recently observed flow anomaly as the forecasted anomaly. Factors to be considered in the generation of persistence forecasts include the duration of the predictor period (i.e., the time span over which the "most recent" flow anomaly is calculated) and the forecast horizon (i.e., the time interval for which this forecast is made). In an analysis of persistence forecast skill for the UK, Svensson (2016) found statistically significant correlations between hindcasts and observations for 78% of station–month combinations using a 1-month predictor period and a 1-month forecast horizon.

Foran Quinn et al. (2021) evaluated the monthly to seasonal hydrological forecast skill of flow persistence for Irish catchments and investigated where and when persistence offers skill beyond climatology using forecast horizons from 1 week to 3 months and which catchment characteristics determine when and where persistence skill is greatest. Forty-six river flow records were selected from the national hydrometric register using the broad criteria adopted by Murphy *et al*. (2013) when creating the Irish Reference Network. A wide range of physical attributes were specified for each catchment by the OPW as part of the FSU (Mills *et al*. 2014). To implement persistence-based forecasts the river flow anomaly observed over a specific "predictor period" (e.g., the month of January) is used as the predicted anomaly for the immediately following "forecast horizon" period (e.g., the month of February). Variations in persistence forecast skill were examined with respect to the duration of these predictor and forecast horizon periods. In each case, the same six durations were tested: 1, 2 and 3 weeks; and 1, 2 and 3 months. Each of the six predictor periods was combined with each of the six forecast horizon periods, meaning that 36 predictor–forecast horizon period combinations were tested for each catchment. Using a combination of a 1-week predictor period and a 1-month forecast horizon, for example, the river flow anomaly over the final week of October would be used to predict the anomaly for the entire month of November. For any given predictor–forecast horizon combination, the predictor-based flow anomalies are therefore used as the hindcast series; the observations in the forecast period are used for model evaluation. Hindcast and observed series of standardized flow anomalies were generated using daily river flow series at each station. These data were used to calculate the mean flow experienced over the predictor period or forecast horizon being tested. For example, a hindcast series of mean weekly flow values was created for the 1-week predictor period, while an observation series of mean monthly flow values was created for the 1-month forecast horizon period. If more than 20% of daily observations in a given period were missing, then this block was not used in the analysis. Following Svensson (2016), mean flows were log-transformed to reduce the influence of extreme flows. For each station, log-transformed mean flow values for a given period, p, of each year were used to calculate the long-term climatological mean flow and standard deviation. The period's log-transformed mean flow value was then converted into a standardized flow anomaly to ensure the distribution of standardized flow anomalies for a given predictor or forecast period at a given station has a mean of 0 and a standard deviation of 1. This standardization approach takes the seasonal cycle of flows into account and enables the comparison of different hydrological regimes. Predicted anomalies can be converted back into predicted flows by reversing the standardization.

The performance of persistence forecasts produced from each predictor–forecast combination was evaluated using Pearson's correlation coefficient (r). The relative predictive performance of the persistence forecasts was then evaluated against the performance of the streamflow climatology benchmark forecasts. The mean squared error skill score (MSESS) was used to assess the improvement (or lack thereof) of the persistence method over climatology in each case. MSESS values range from $-\infty$ (least skilful) to 1 (perfectly skilful), with any positive value representing skill relative to the benchmark (climatology). The relationship between available physical catchment descriptors and the predictive skill of river flow persistence was evaluated using Spearman's rank correlation (ρ) to help explain why flow anomalies tend to persist more in some catchments than others. An all-subsets regression approach was used in conjunction with cross-validation to find the best multivariate regression models at predicting catchment annual and seasonal average persistence skill.

Across all catchments and predictor periods, persistence skill declines with increasing forecast horizon. Figure 1 shows the decay in network-wide correlations between hindcast and observed anomalies for the 1-week forecast (r = 0.69) through to the 3-month ahead forecast (r = 0.31) when holding the predictor period at 1 week. The upper (95th percentile) and lower (5th percentile) bands show considerable variation in forecast performance at each horizon across the station network. For example, the 1-month forecast horizon has mean r = 0.52 with 95th percentile range spanning r = 0.18 to r = 0.81. Similarly, across all catchments and forecast horizons, annual average persistence skill declines as the duration of the predictor period increases. Figure 1 shows a gradual decrease in network-wide correlations between hindcast and observed anomalies for the 1-week predictor period (r = 0.52) through to the 3-month predictor period (r = 0.24), when holding the forecast horizon at 1 month. Again, there is significant variation about the mean depending on catchment. Overall, the persistence forecasts perform best using the 1-week predictor period and 1-week forecast horizon. However, as we focus on monthly to seasonal hydrological forecasting, this 1-week predictor period was combined with 1- to 3-month forecast horizons.

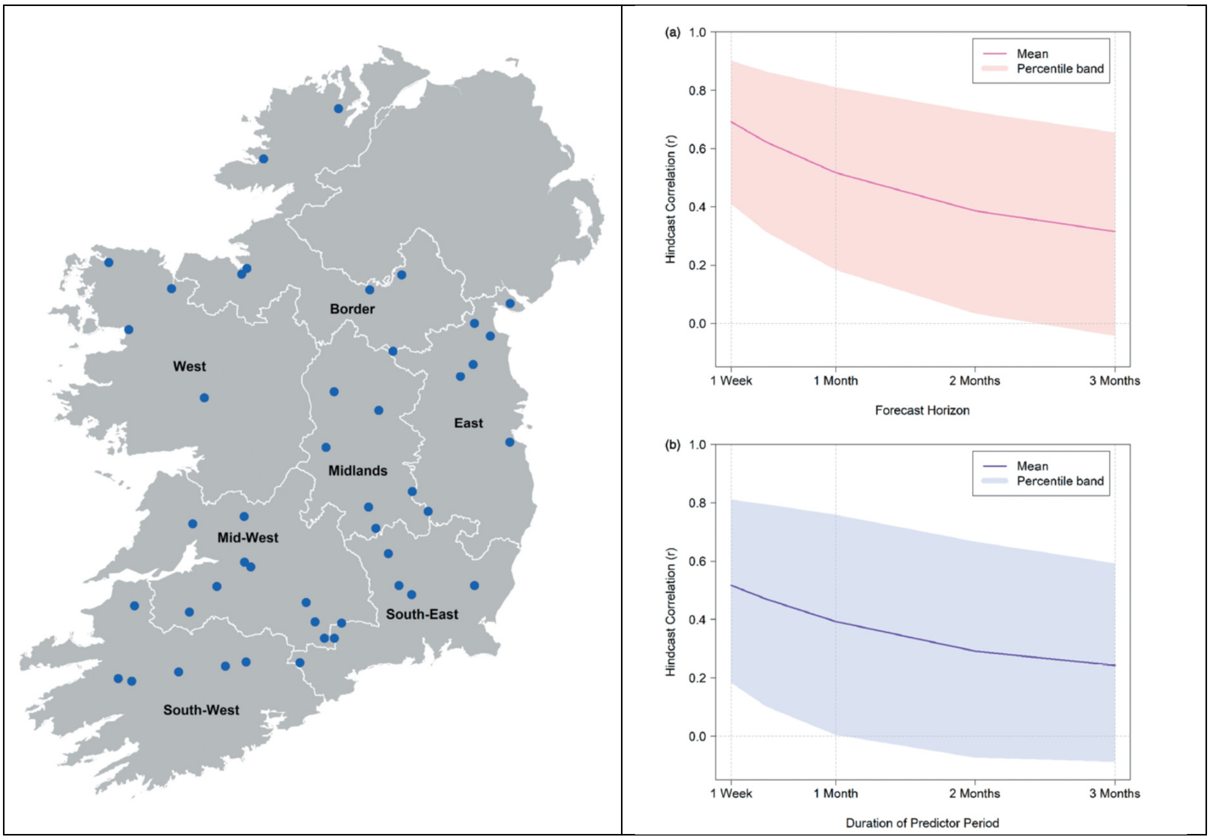

Figure 1: Left: location of the 46 stations (blue dots) used for the evaluation of persistence forecasts and regions in the Republic of Ireland, based on the Nomenclature of Territorial Units for Statistics (NUTS) Level III classification. Right: Network-wide persistence forecast performance, measured by the correlation (*r*) between hindcasts and observations, plotted against: (a) forecast horizon, using a 1-week predictor period, and (b) duration of the predictor period, using a 1-month forecast horizon. The spread of persistence skill across the network is indicated by the upper (95th percentile) and lower (5th percentile) bands.

Performance of the flow persistence forecasts varies throughout the year. A distinct seasonal pattern can be identified in forecast performance, with summer months (JJA) having the highest seasonal mean correlation coefficient (*r* = 0.66) and winter months (DJF) having the lowest (*r* = 0.44). Spring (MAM) and autumn (SON) have similar mean correlation values of *r* = 0.49 and 0.48, respectively. This seasonality of forecast performance becomes more pronounced as longer predictor periods and/or forecast horizons are used. The majority (58%) of persistence forecasts produced across the catchment sample perform better than the streamflow climatology benchmark at the 1-month horizon (Fig. 2). The following qualitative descriptors are used to categorize MSESS values as high (0.5–1), moderate (0.25–0.5), low (0–0.25) and no skill (≤ 0).

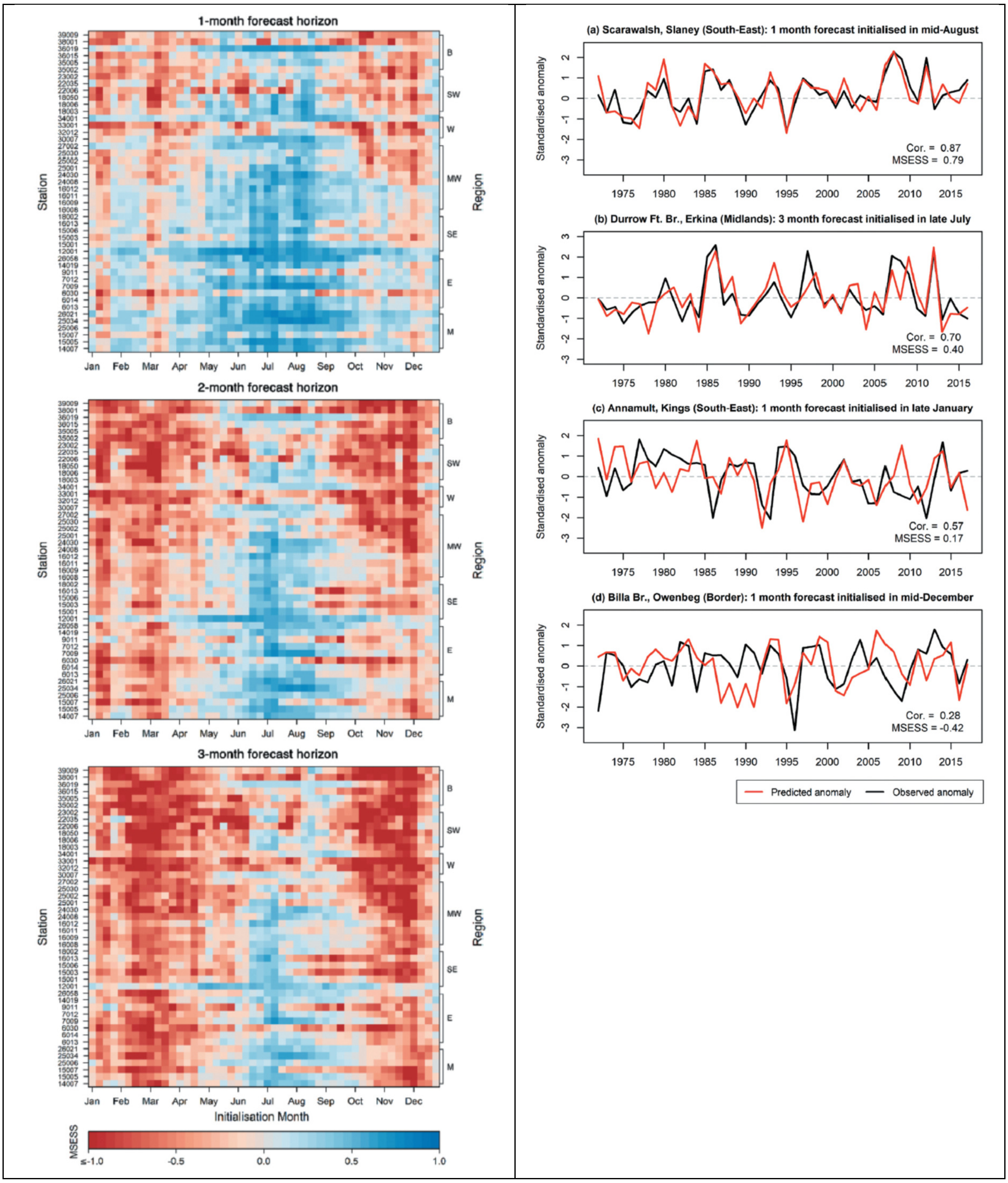

Figure 2: Left: Variation in forecast skill based on a 1-week predictor period and 1- to 3-month forecast horizons across all weekly initializations for each station, grouped by their respective regions: B (Border), SW

(South-West), W (West), MW (Mid-West), SE (South-East), E (East) and M (Midlands). Right: Hindcast time series comparing the observed (black) and predicted river flow anomaly using the persistence method (red) at sample forecast horizons for four example stations at four example initialization times. The examples show skill scores (MSESS) that are: (a) high, (b) moderate, (c) low and (d) negative.

Figure 2 also presents four example hindcast time series to illustrate variations in skill across these different categories. The proportion of skilful forecasts is higher for forecasts initialized during the summer months (87%), compared with spring and autumn (both 53%). Winter is the least skilful season, with only 41% of simulations outperforming climatology. The most skilful predictor month is August. March has the lowest average skill score. Stations with the best-performing persistence forecasts are mainly found in the Midlands, East and South-East regions – collectively, these have median $r = 0.59$ between hindcasts and observations. Conversely, significantly lower median $r = 0.41$ is found between hindcast and observed anomalies in the catchments of the Border, West and South-West regions.

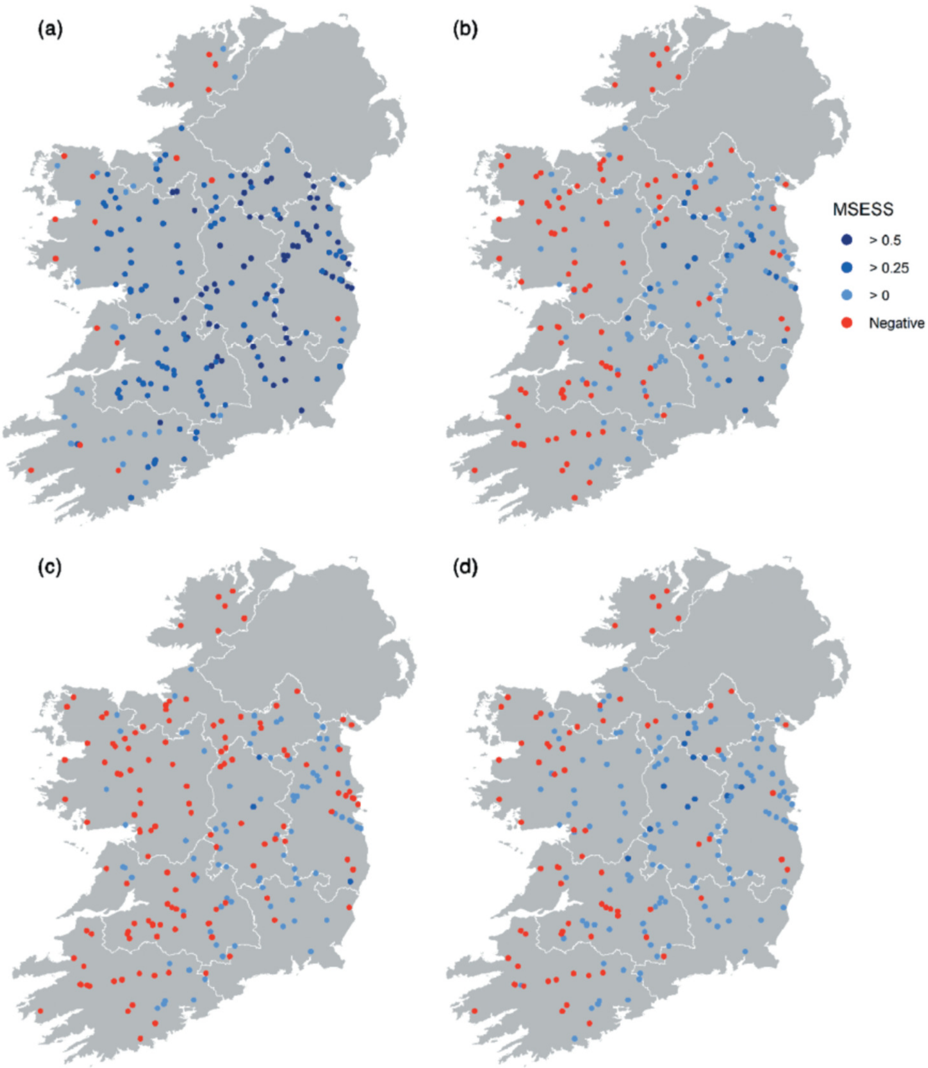

Figure 3: Median predicted seasonal persistence skill scores (MSESS) for 215 FSU catchments in: (a) summer, (b) autumn, (c) winter and (d) spring based on a 1-week predictor period and 1-month forecast horizon.

A strong positive correlation is seen between median persistence skill and BFIsoil ($\rho = 0.86$), indicating that persistence skill is greater for catchments with higher storage capacities. Skill is moderately negatively correlated with SAAR ($\rho = -0.66$) and other physical descriptors linked to the wetness of a catchment. The best-performing multiple linear regression model used BFIsoil, SAAR, SAAPE, S1085 and TAYSLO as predictors of annual median catchment persistence skill for the 1-week predictor period and 1-month forecast horizon, yielding an adjusted R2 of 0.90. The regression models that provided the most accurate predictions of the sample catchments' median seasonal MSESS were used to infer the likely average persistence skill of the larger FSU 215-catchment set (Fig. 3). This provides an overview of the expected performance of the flow persistence method outside the training set, allowing exploration of potential utility of the technique as an operational forecasting tool at the national scale. These results could also be used to inform strategic investment in hydrometric networks by identifying where gauges could be most successfully used for persistence-based forecasts.

These results show that river flow persistence offers a source of skill for monthly to seasonal hydrological forecasting in Irish catchments. This skill is conditional on the duration of the predictor and forecast horizon periods, the forecast initialization month and catchment characteristics. Using the most skilful predictor period (1 week), most persistence forecasts outperform the river flow climatology benchmark in February and from April to September at the 1-month horizon. However, this narrows to the summer months when using 2- and 3-month forecast horizons. This reduction in skill results from the longer time a catchment has to "forget" initial anomalous river flow conditions and/or to be impacted by "new" anomalies. Predictability relies on catchment memory, and hence skill is strongly positively correlated with BFIsoil ($\rho = 0.86$). As such, persistence forecast skill is greatest in lowland regions that are characterized by permeable lithologies, well-drained soils and lower annual average rainfall totals. The seasonal and spatial distribution of persistence skill shows when and where initial hydrological conditions provide useful seasonal flow forecasting skill (as in the slowly responding catchments predominantly located in the Midlands, east and south-east of the island). It also underscores the scope for development of dynamical hydrological forecasting approaches in the wetter, poorly drained catchments underlain by impermeable lithologies, found mainly in the Border, West and South-West regions.

3. ENSEMBLE STREAMFLOW PREDICTION (ESP) AND CONDITIONING ON NAO FORECASTS FROM DYNAMICAL MODELS

We 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. The conventional ESP approach 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, and streamflow itself) (Wood and Lettenmaier, 2008). ESP is recognised as a low-cost, "tough-to-beat" forecast (Pappenberger et al., 2015) against which value added by more sophisticated hydrometeorological ensemble systems can be assessed (e.g., Arnal et al., 2018; Wanders et al., 2019). Hence, the potential application of ESP in Ireland merits exploration.

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. 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), 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.
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. Reliability and discrimination of forecasts are assessed with respect to low- and high-flow events. 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 (x1–x4) that require calibration of observed streamflow data against precipitation and potential evaporation. We find the model 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). See Donegan et al. (2021) for details of model calibration.

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 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 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. Hindcast time series were therefore temporally aggregated to provide predictions of mean streamflow over lead times of 1 d to 12 months, resulting in 365 lead times per forecast (excluding leap days). 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 were not used as input to the model, as these 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 d in length (1 January to 31 December) extracted from 1961–2016 but not 1965. ESP skill is evaluated over 52 initialisation years (1965–2016) with 12 initialisation months (January to December).

To investigate the potential for improving winter streamflow predictability, we conditioned the ESP method above 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 (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 (a) to 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 2009–2010 if below −10 hPa (Stringer et al., 2020). Per hindcast member, 10 analogues were sampled to minimise non-NAO-related variability whilst keeping a consistent NAO signal across the sample. Conditioned ESP forecasts were only initialised on 1 December.

We quantify the skill of the ESP method using the continuous ranked probability score (CRPS; Hersbach, 2000) and corresponding skill score (CRPSS). The CRPS is a recommended and widely used evaluation metric for ensemble hydrological forecasting (Pappenberger et al., 2015) 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.

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.

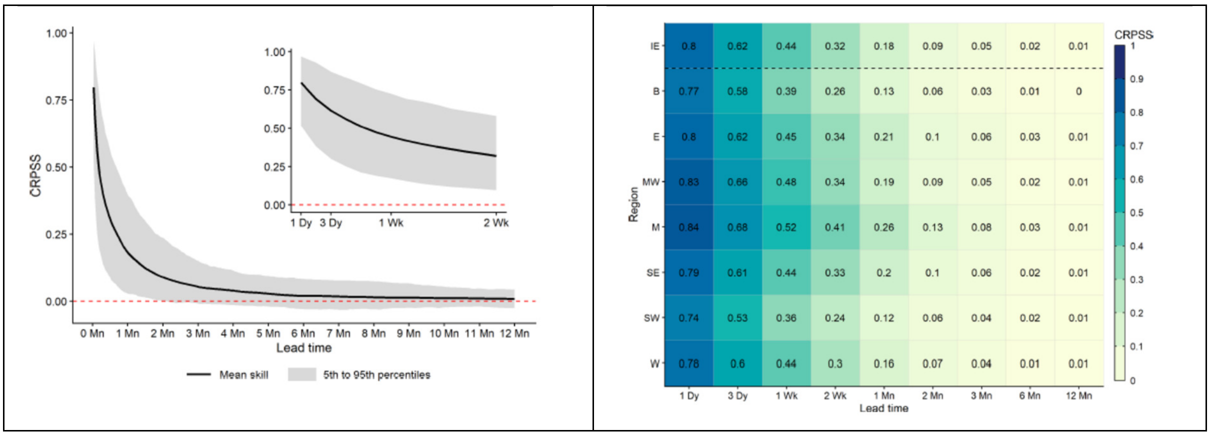

Figure 4: Left: Mean ESP CRPSS values across all 46 study catchments, 12 forecast initialisation months, and all 365 lead times, with short and extended lead times shown inset for readability. Variations in skill scores across all catchments at each lead time are given by the 5th and 95th percentile ensemble range. Right: 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 d), extended (1- and 2-week), monthly (1- and 2-month), seasonal (3- and 6-month) and annual (12-month).

Mean ESP skill declines rapidly as a function of lead time, across all catchments and initialisation months (Fig. 4). Mean CRPSS values for short (1 d) 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 5th and 95th percentile bands. ESP skill varies with forecast initialisation month and time of year, with the 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). The Midlands, Mid-West, and East are the most skilful regions, followed by the South-East, West, and Border regions (Fig. 4). 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. As with the persistence method above, differences in regional hydroclimate properties 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.

ESP is skilful at forecasting the occurrence of both low-flow (lower tercile) and high-flow (upper tercile) events up to 1 month ahead in the majority of catchments and for all initialisation months (Fig. 5). 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. 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.

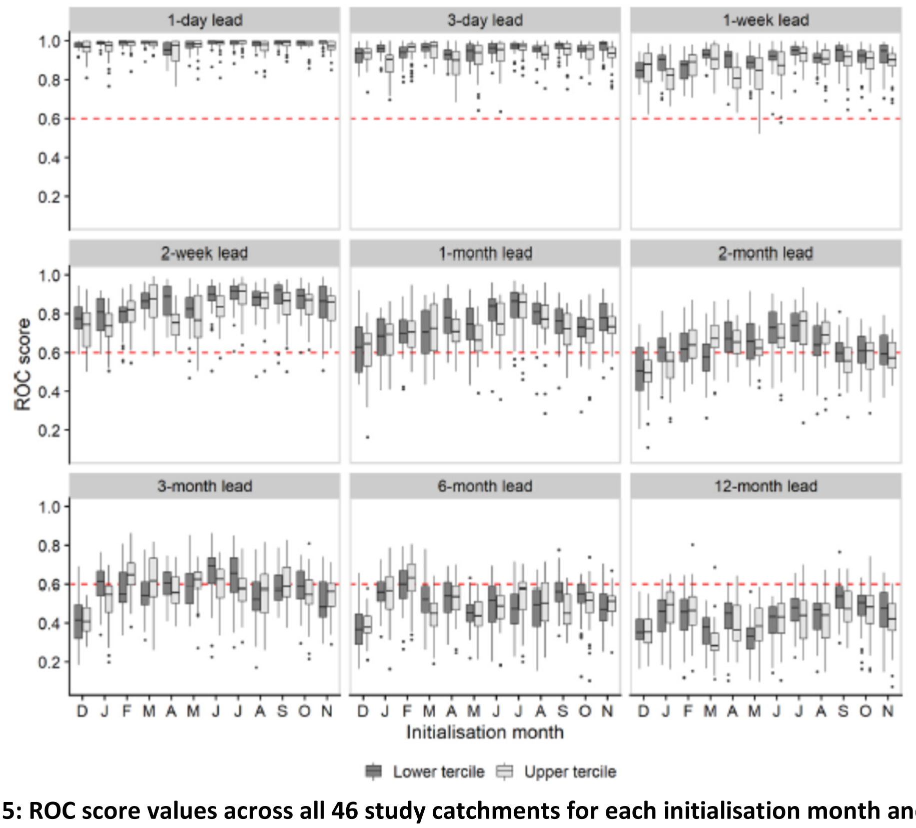

Figure 5: ROC score values across all 46 study catchments for each initialisation month and the same selection of lead time

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. At a lead time of 2 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, the greatest improvements in NAO-conditioned ESP are found for wet, fast-responding catchments with low baseflow contribution.

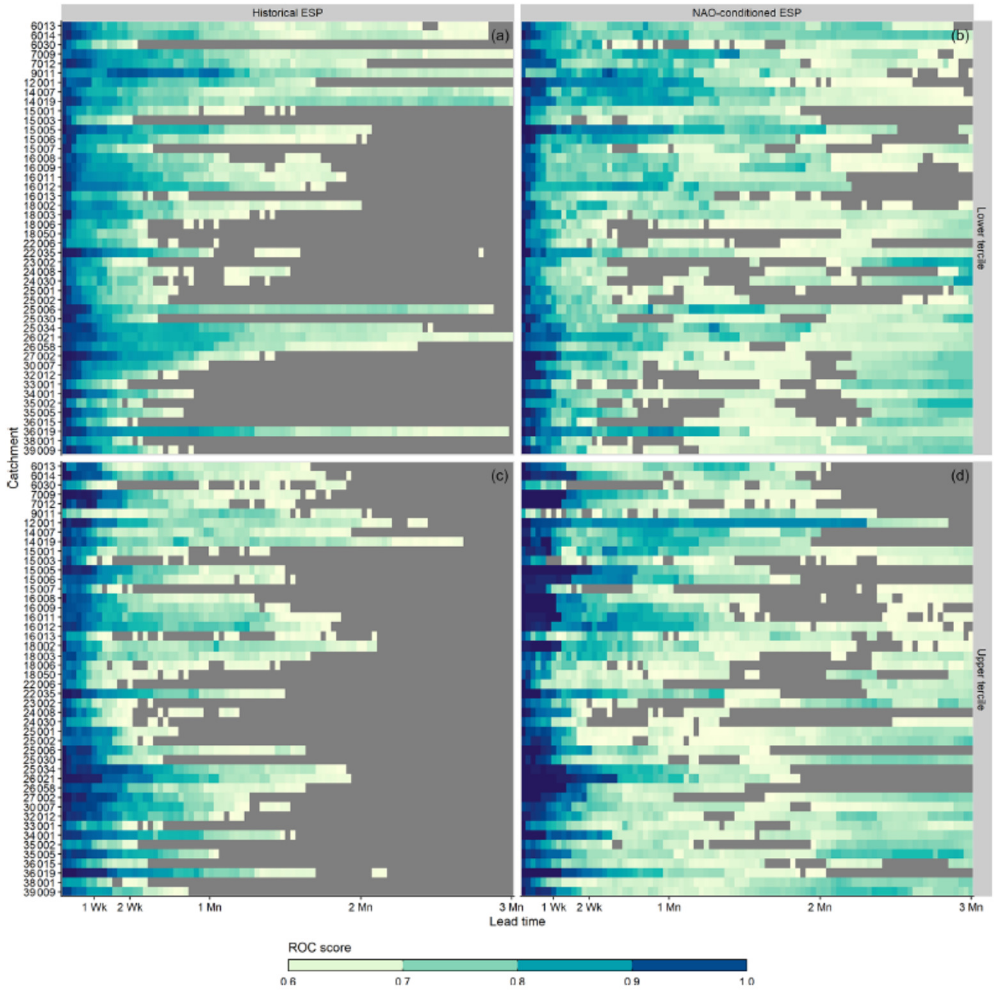

Figure 6: Comparison of ROC scores achieved by historical ESP (a, c) and NAO-conditioned ESP (b, d) across all 46 study catchments and all lead times for low-flow (lower tercile, a, b) and high-flow (upper tercile, c, d) events. Cells with no skill (ROC < 0.6) are greyed out.

ROC scores for individual catchments and the full range of lead times are presented in Fig. 6. 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 to 89 d) and 170 % for high flows (33 to 89 d). 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 is 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. 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. DYNAMICAL-STATISTICAL SEASONAL FORECASTS OF WINTER AND SUMMER PRECIPITATION

Seasonal precipitation forecasting is highly challenging for the northwest fringes of Europe due to complex dynamical drivers. Hybrid dynamical–statistical approaches offer potential to improve forecast skill. Here, hindcasts of mean sea level pressure (MSLP) from two state of the art dynamical seasonal forecasting systems (GloSea5 and SEAS5) were used to derive two distinct sets of indices for forecasting winter (DJF) and summer (JJA) precipitation over lead-times (LT) of 1–4 months. These indices provide predictors of seasonal precipitation via a multiple linear regression model (MLR) and an artificial neural network (ANN) applied to four Irish rainfall regions and the Island of Ireland (see Golian et al., 2022 for details).

Two predictors were derived from MSLP hindcasts. First, correlation maps for seasonal mean MSLP from ERA5 versus observed precipitation were first generated for each precipitation region to identify locations of maximum and minimum correlation. We used MSLP for the domain 50°W–50°E and 20°–80°N. Like Baker *et al*. (2018) fixed points of maximum and minimum correlation between MSLP and precipitation from observations are used to derive the standardized MSLP index from model hindcasts for each precipitation region. Depending on the source of the MSLP hindcasts (i.e., GloSea5 or SEAS5), the index is defined as the standardized (i.e., centred about the mean and divided by the standard deviation over the time series) MSLP difference between the fixed maximum and minimum points of correlation from ERA5 MSLP. This index was derived for each precipitation region and lead time separately using the ensemble mean.

Second, we applied a rotated empirical orthogonal function (REOF) as in Hall and Hanna (2018). ROEF analysis can avoid artificial dipole-type patterns which can be produced by traditional EOF analysis (Lian and Chen, 2012). Application of REOF to MSLP anomalies was undertaken separately for winter and summer with respect to the long-term seasonal mean (1993–2016 for GloSea5; 1994–2016 for SEAS5). To account for latitudinal variation in grid cell areas, MSLP anomalies were weighted by the cosine of the latitude prior to analysis. The three leading vectors of the cross-correlation matrix calculated from monthly MSLP from GloSea5 and SEAS5 hindcasts were used to construct the set of indices from each dynamical forecasting system. Forecast skill for each model, lead time, and region was evaluated using the correlation coefficient (*r*) and mean absolute error (MAE), benchmarked against (a) climatology, (b) bias corrected precipitation hindcasts from both GloSea5 and SEAS5, and (c) a zero-order forecast based on rainfall persistence.

Hindcasts for winter and summer precipitation with 1-month lead-time are presented in Figures 7 and 8. Overall, for both seasons, regression and ANN hindcasts outperform the persistence, climatology, and bias-corrected dynamical model output for all regions. Moreover, at the 1-month lead-time, the ANN and MLR show some skill at predicting extreme seasons. For example, the dry summer of 1995 and wettest winter on record 2015/2016 are captured well by our hybrid models. Scaife *et al*. (2017) show that the intensified cyclonic flow over the Atlantic in winter 2015/2016 were well predicted by the GloSea5 system. However, the extremeness of the wet 2013/2014 winter and sequence of exceptionally wet summers in 2007–2009 are underestimated (Matthews *et al*., 2014; 2016). Knight *et* *al.* (2017) assert that the tropics played a significant role in the development of the unusual extratropical circulation that led to widespread high precipitation over the UK in winter 2013–2014.

In summary, our new dynamical–statistical methods perform satisfactorily in prediction of precipitation up to 4 months ahead, surpassing all available benchmarks in both winter and summer. The ANN performs better than MLR in most regions in summer especially for LT1 and LT3, and in winter for LT2, LT3 and LT4. However, the MLR marginally outperforms ANN in most regions at LT1 in winter and LT4 in summer. The skill of bias-corrected precipitation forecasts from SEAS5 and GloSea5 are generally not as good as climatology in winter and only marginally better than persistence in summer for some lead-times/regions.

Among the potential predictors available, those based on standardized MSLP showed strongest correlations (spanning 0.35–0.63) with precipitation across both seasons, all lead times and regions. This highlights that indices based on MSLP provide a reliable basis for forecasting precipitation over the Island of Ireland, especially in winter. MLSP might also enhance seasonal forecast skill for other climate variables and regions where there is covariance with precipitation, such as for atmospheric humidity, air temperature, sunshine hours and wind speeds (e.g., Hillier *et al*., 2020).

Figure 7: Winter [DJF] precipitation hindcasted for each region for LT1. Results are shown for EOBS observations (black line), the MLR (green), the ANN (blue), persistence (grey dashed), bias corrected GloSea5 (orange), SEAS5 (red) and climatology (grey) precipitation.

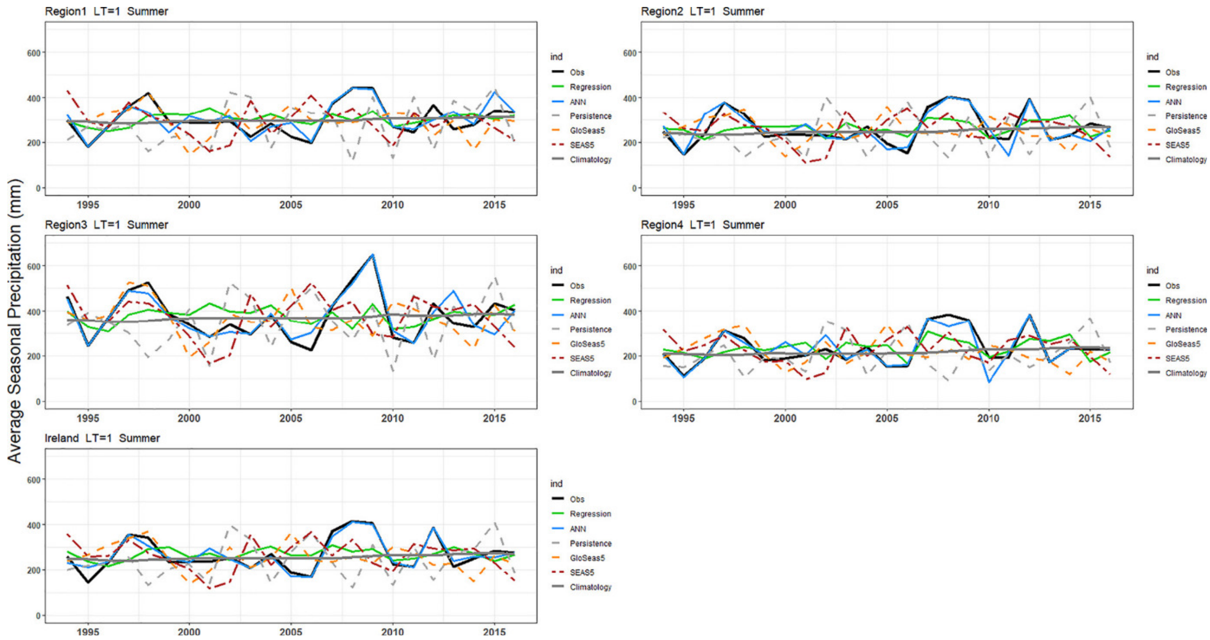

Figure 8: As Fig. 7 but for summer [JJA].

Interestingly, some of the strongest correlations were found between summer precipitation and EOF3 from GloSea5 data, with this index selected as a predictor for many regions/lead-times. This EOF is usually interpreted as the East Atlantic (EA) pattern (Hall and Hanna, 2018; Comas-Bru and McDermott, 2014). Hence, although the NAO has long been recognized as a major driver of climate variability in Northwest Europe, we note that an apparent EA pattern emerges as a key signal of summertime predictability in the dynamical models.

Our findings demonstrate the feasibility of skilful seasonal forecasts of winter and summer precipitation in Ireland. Such forecasts could be of value to many sectors, not least the water sector. The dynamical–statistical approach developed here leverages two forecasting systems to develop prototype predictions of winter and summer precipitation with up to 4 months lead for regions across the Island of Ireland. Our ANN and regression models show consistently better results compared with bias corrected dynamical outputs, climatology, and persistence methods. Although our dynamical–statistical models performed well in both winter and summer, the ANN provided marginally higher skill in winter, and the MLR in summer. The encouraging performance of our models, particularly in summer, is noteworthy and paves the way to improved drought forecasting in summer and winter. However, before evaluating their operational utility, future work should extend the analysis of uncertainty. For example, individual ensemble members rather than just the ensemble mean could be used to generate probabilistic seasonal forecasts. There is also the possibility of assessing other candidate predictors and domains in future work. For instance, bias-corrected precipitation forecasts from dynamical systems could be used as inputs to data-driven models alongside the MSLP-based indices employed here.

5. CONCLUSIONS

Results from the HydroCast project to date evidence the potential for deployment of SHF methods to support the water sector. Persistence based methods offer skill for predicting river flow anomalies by leveraging initial hydrological conditions, especially when initialised during summer months in catchments in the southeast and midlands where groundwater storage is present. Results from Ireland show that skill from persistence-based methods are comparable to those for the UK where this method is routinely used as part of the Hydrological Outlook. ESP methods add complexity due to introduction of rainfall runoff models. We show that predictions of winter NAO conditions from the UK Met Office GloSea5 model can offer substantially improved skill in catchments where predictive skill from initial hydrological conditions is low (catchments in west and northwest) and adds value in the ability to predict dry winter conditions, which is important for water management. Finally, we have shown that while direct use of precipitation outputs from dynamical forecasting systems offers limited skill, we can leverage predictions of sea level pressure together with statistical/machine learning methods to develop useful forecasts of summer and winter precipitation at lead times of up to four months. This skill is likely to be extendable to other variables that are correlated with SLP. These advances in understanding and methods show when and where SHF can offer utility to water management. Ongoing research in the project is evaluating the added skill from ESP methods that can be realised from using more complex rainfall runoff models, the utility of bias correction methods for leveraging increased skill from ensemble precipitation forecasts from dynamical systems and the ability to develop predictive models of low flows from key modes of climate variability affecting the region.

REFERENCES

Arnal, L. et al. (2018) Skilful seasonal forecasts of streamflow over Europe?, *Hydrology and Earth System Sciences*, 22, 2057–2072.

Baker, L. H. et al. (2018) Improved seasonal prediction of UK regional precipitation using atmospheric circulation, *International Journal of Climatology*, 38, e437–e453.

Bell, V. et al. (2017) A national-scale seasonal hydrological forecast system: development and evaluation over Britain, *Hydrology and Earth System Sciences*, 21, 4681–4691.

Bierkens, M. F. P. and van Beek, L. P. H. (2009)Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, *Journal of Hydrometeorology*, 10, 953–968.

Comas-Bru, L. and McDermott, F. (2014) Impacts of the EA and SCA patterns on the European twentieth century NAO–winter climate relationship. *Quarterly Journal of the Royal Meteorological Society*, 140(679), 354– 363.

Demargne, J. et al. (2010) Diagnostic verification of hydrometeorological and hydrologic ensembles, *Atmospheric Sciences Letters*, 11, 114–122.

Donegan, S. et al. (2021) Conditioning ensemble streamflow prediction with the North Atlantic Oscillation improves skill at longer lead times. *Hydrology and Earth System Sciences*, 25(7), 4159-4183.

Ferro, C. A. T. et al. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores, *Meteorological Applications*, 15, 19–24.

Foran Quinn, D. et al (2021) Benchmarking seasonal forecasting skill using river flow persistence in Irish catchments. *Hydrological Sciences Journal*, 66(4), 672-688.

Golian, S. et al. (2022) Dynamical–statistical seasonal forecasts of winter and summer precipitation for the Island of Ireland. *International Journal of Climatology*, https://doi.org/10.1002/joc.7557

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

Harrigan, S. et al. (2018) Benchmarking ensemble streamflow prediction skill in the UK, *Hydrology and Earth System Sciences*, 22, 2023–2039.

Hersbach, H. (2000) Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather Forecasting*, 15, 559–57.

Hillier, J.K. et al. (2020) Multi-hazard dependencies can increase or decrease risk. *Nature Climate Change*, 10(7), 595– 598.

Knight, J.R. et al. (2017) Global meteorological influences on the record UK rainfall of winter 2013–2014. *Environmental Research Letters*, 12(7), 074001.

Lian, T. and Chen, D. (2012) An evaluation of rotated EOF analysis and its application to tropical Pacific SST variability. *Journal of Climate*, 25(15), 5361– 5373.

MacLachlan, C. (2015) Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system, *Quarterly Journal of the Royal Meteorological Society*., 141, 1072–1084.

Mason, S. J. and Graham, N. E. (1999) Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecasting*, 14, 713–725.

Matthews, T. et al. (2014) Stormiest winter on record for Ireland and UK. *Nature Climate Change*, 4(9), 738– 740.

Matthews, T. et al. (2016) Past and future climate change in the context of memorable seasonal extremes. *Climate Risk Management*, 11, 37– 52.

Mills, P. et al. (2014) Physical catchment descriptors. Volume IV, Flood studies update technical research report. Dublin: Office of Public Works.

Murphy, et al. (2013) HydroDetect: the identification and assessment of climate change indicators for an Irish reference network of river flow stations. Wexford: Environmental Protection Agency.

Pappenberger, F. et al. (2015) How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction, *Journal of Hydrology*, 522, 697–713.

Perrin, C. et al. (2003) Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, 279, 275– 289.

Prudhomme, C., et al. (2017) Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal timescales. *Hydrological Sciences Journal*, 62 (16), 2753–2768.

Scaife, A. A. et al. (2014) Skillful long-range prediction of European and North American winters, *Geophysical Research Letters*, 41, 2514–2519.

Scaife, A.A. et al. (2017) Predictability of European winter 2015/2016. *Atmospheric Science Letters*, 18(2), 38– 44.

Smith, D. M. et al. (2020) North Atlantic climate far more predictable than models imply, *Nature*, 583, 796– 800.

Steirou, E. et al. (2017) Links between large-scale circulation patterns and streamflow in Central Europe: A review, *Journal of Hydrology*, 549, 484–500.

Stringer, N. et al. (2020) Improving Meteorological Seasonal Forecasts for Hydrological Modeling in European Winter, *Journal of Applied Meteorology and Climatology*, 59, 317–332.

Svensson, C. (2016) Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues. *Hydrological Sciences Journal*, 61 (1), 19–35.

Svensson, C. et al. (2015) Long-range forecasts of UK winter hydrology, *Environmental Research Letters*, 10, 064006.

Wanders, N. et al. (2019) Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System, *Journal of Hydrometeorology*, 20, 99–115.

Webster, K.E., et al. (2014) Environmental flow assessment for Irish Rivers. Wexford: Environmental Protection Agency.

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

Wood, A. W. and Lettenmaier, D. P. (2008) An ensemble approach for attribution of hydrologic prediction uncertainty, *Geophysical Research Letters*, 35, L14401.

Wood, A. W. et al. (2016) Quantifying Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill, *Journal of Hydrometeorology*, 17, 651–668.