



Supplement of

Improved seasonal hydrological forecasting for Great Britain

Mark D. Rhodes-Smith et al.

Correspondence to: Mark D. Rhodes-Smith (marrho@ceh.ac.uk)

The copyright of individual parts of the supplement might differ from the article licence.

Supplemental Material to Rhodes-Smith et al. (2026) *Improved seasonal hydrological forecasting for Great Britain*

S1. Construction of Historic Weather Analogues rainfall forecasts

For each forecast period, the Met Office's Global Seasonal forecasting model version 6 (GloSea6; Met Office, 2024a) is first used to generate an ensemble of seasonal (monthly) Mean Sea Level Pressure (MSLP) fields covering the North Atlantic and Europe. This MSLP ensemble represents our understanding of, and the uncertainty in, the potential evolution of contemporary atmospheric conditions over the season (month). The Met Office initialises four runs of GloSea6 each day, two 'short' runs of 64 days and two 'long' runs of 216 days. The seasonal (monthly) forecast ensembles are constructed by aggregating these runs from multiple initialisation days. For the monthly forecast, the four members (two long and two short) are aggregated over the preceding 11 initialisation days, giving an ensemble of 44 members. For the seasonal forecast, only the two long members can be used. These are aggregated over the preceding 21 initialisation days, giving an ensemble of 42 members. The monthly and seasonal forecasts are issued in the week before the start of the forecast period, and are normally prepared between the 20th and 25th day of the month. These GloSea6 ensemble members are thus derived from atmospheric conditions preceding the initialisation of the hydrological forecasts by at least five and up to thirty-two days.

We next select the rainfall patterns that correspond to these MSLP forecasts. We will do this by identifying the years in which the observed MSLP pattern over the same temporal and spatial domain is most similar to the forecast field, from which we will adopt the observed rainfall as our forecast. However, the number of historic observations is small (only 63 years from 1961 to 2023 inclusive), and in consequence are a limited sample of potential MSLP fields.

We therefore construct an array of synthetic MSLP fields from combinations of the historical observations and make our selection from these. For a seasonal forecast, the synthetic fields are derived from the seasonal mean of all combinations of the corresponding three months in the historical record. For instance, for a summer forecast one such combination is the average of the monthly MSLP fields from June 1971, July 1972 and August 1973. For a one-month forecast, synthetic months are constructed in a similar manner by averaging MSLP fields from combinations of three sub-monthly periods consisting of the first 10 days, second 10 days, and the remainder of the month. For instance, a synthetic July MSLP field can be constructed from the combination of 1st-10th July 1971, 11th-20th July 1972, and 21st-31st July 1973. In both cases, our synthetic array consists of all combinations of years between 1961 and 2023 (inclusive, yielding 63 years), resulting in $63^3=250047$ synthetic MSLP fields. This number is sufficiently large as to allow the selection of an ensemble of analogues with minimal duplication and spans the range of possible forecasts (better than the observations; see Stringer et al., 2020). Using more than three analogues within each synthetic season/month would increase the computational cost and introduce additional discontinuities into the forecasts as we transition between analogues.

For each GloSea6 ensemble member, we calculate the residual at each grid point between the ensemble member's MSLP anomaly field and the MSLP anomaly field for each synthetic season (month). We then calculate the area-weighted root-mean-square of these residuals over the full spatial domain, yielding a scalar value for each synthetic season (month). The 10 synthetic seasons (months) with the smallest values are chosen as 'resamples' of the GloSea6 member.

The use of more than one synthetic field for each GloSea6 member was proposed by Stringer et al. (2020) to account for the range of rainfall patterns that can arise from any given MSLP field. An additional benefit is that it increases the ensemble size, resulting in a smoother, better sampled, forecast probability distribution. This is particularly valuable in the tails of the distribution, where extreme events are found. To achieve the same ensemble size by running the GloSea6 model an increased number of times would be prohibitively expensive.

For seasonal forecasts between November-December-January (NDJ) and February-March-April (FMA), the selection of analogue years differs slightly. At this time of the year, the “signal-to-noise paradox” is most evident, whereby the skill of a seasonal forecast is greater than can be explained by the relatively modest size of predicted signals, leading to under-confident forecasts (Eade et al., 2014; Scaife and Smith, 2018). This has a significant implication for the ultimate hydrological forecasts, since if the signal size is small then there are only ever small shifts in the probability density function of the input rainfall, regardless of how good the skill is. In other words, this will lead to frequent underestimation of the magnitude of deviations in streamflow from normal. To account for the effects of the signal-to-noise paradox, the magnitude of the predicted circulation signal is inflated prior to the selection of the analogues for the NDJ to FMA forecasts (as in Stringer et al. 2020). This is done as follows:

- (1) We first calculate the North Atlantic Oscillation (NAO) indices for each of the three monthly-mean MSLP fields of each GloSea6 ensemble member. These are adjusted such that the ensemble mean of the seasonal-mean NAO index is inflated by a factor to account for the magnitude of the signal-to-noise paradox in that season, but such that the total variance (the combination of the long-term variance in the ensemble means over many forecasts and the variance due to noise in the particular ensemble) is preserved. The derivation of this factor and the required adjustments to the NAO index of each ensemble member are provided in Sect. 4b of Stringer et al. (2020).
- (2) A set of 100 synthetic seasons is then constructed by randomly selecting (with replacement) for each month analogue years in which the monthly-mean NAO index is within 4hPa of the adjusted GloSea6 ensemble member. This emphasises the NAO signal compared to the set of all combinations of possible analogue years used in the rest of the year.
- (3) For each synthetic season in our 100-member subsample we calculate the MSLP residuals by subtracting out the component of the MSLP field corresponding to the NAO index. This contribution is estimated at each grid point from a linear regression between historic NAO indices and MSLP fields. Similarly, we calculate the MSLP residuals for the adjusted GloSea6 member. We finally select the 10 unique synthetic seasons from our 100-member subsample that minimise the area-weighted root-mean-square difference between these MSLP residuals.

The result of the process so far is an ensemble of analogue years (with each ensemble member consisting of the combination of three years) from which rainfall forecasts (and in principle other meteorological predictands) can be constructed as follows. We first construct an ensemble of daily rainfall sequences for the forecast period using observations taken in the selected combinations of analogue years. In this work, we derive these from the 1km HadUK-Grid dataset (v1.3.0.ceda; Hollis et al., 2019; Met Office, 2024) which intrinsically matches the resolution of the hydrological model. For a seasonal forecast we obtain the observed daily total rainfall from each of the three months in the analogue years. For a monthly forecast, the observed daily total rainfall in the periods 1st-10th, 11th-20th and the remainder of the month in the three analogue years are used. These daily sequences are then summed to obtain the forecast monthly total rainfall.

A ‘detrending shift’ is then applied to the forecasts to correct for the long-term trends in rainfall between the analogue year and the forecast year. This shift differs in each month and is estimated as follows, here using January as an example:

- (1) We first construct a dataset consisting of the averages of the monthly-mean NAO indices and monthly-total 1km observed rainfall maps from December, January and February from each year 1961-2022 (inclusive). We include the months adjacent to the one of interest to reduce the noise in the trend estimate.
- (2) We now want to calculate the component of each sample’s rainfall that can be attributed to the NAO. This will be excluded in the calculation of the detrending factor so that our resulting forecasts retain sensitivity to the long-term trends in the NAO. To do this, a high-pass filter is applied to the NAO index and to each pixel of the 1km rainfall grids to estimate the high-frequency (year-to-year) variability. This excludes the long-term (decadal) trends we will ultimately seek to calculate. We then perform a linear regression to the high pass filtered NAO and rainfall. This provides a function that maps the NAO index into the resulting rainfall, allowing us to estimate the part of the rainfall that can be attributed to the NAO. We finally remove the NAO-associated rainfall from the original observed rainfall grids collected in step 1. This leaves us with residual rainfall maps.
- (3) The long-term trends in the rainfall (that do not result from variation in the NAO) are then estimated for each pixel by a linear fit to the time series of rainfall residuals from Step 2.

This detrending shift is then applied to the rainfall forecasts, such that

$$P'_m = P_m + D_m(Y - Y') \quad (1)$$

where P'_m and P_m are the detrended and original rainfall for month m respectively, D_m the corresponding detrending factor and $(Y - Y')$ the difference between the forecast year and the analogue year. If the predicted monthly total rainfall is negative, which can arise when D_m is negative, we set it to zero.

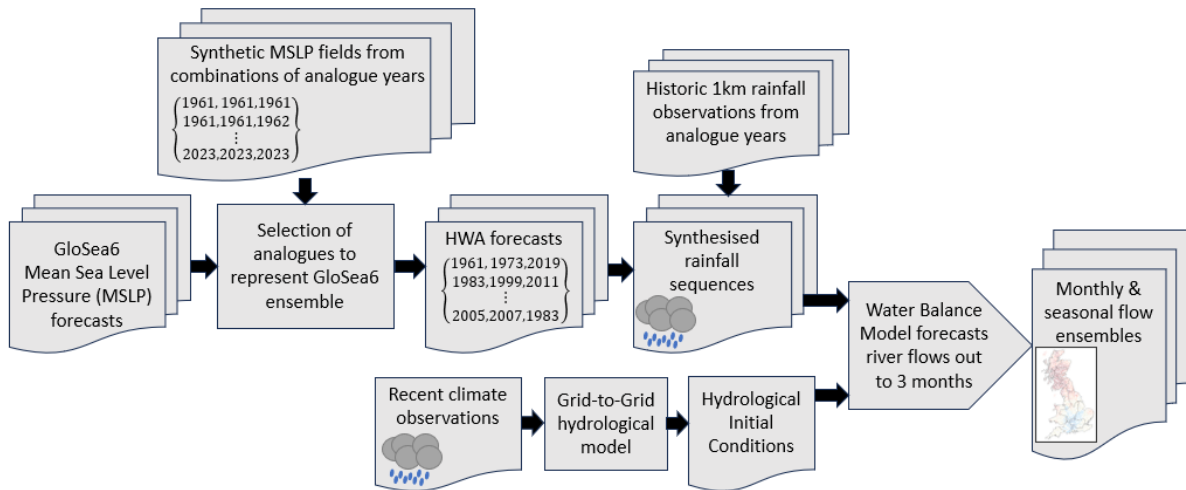


Figure S1: Schematic of the G2G/WBM hydrological forecasting scheme, here driven by rainfall forecasts derived from Historic Weather Analogues.

S2. Performance of the operational scheme

The G2G/WBM seasonal forecasting system has been in commissioning (October 2022 to November 2023) and operational (since December 2023) for a small number of years. This gives us a recent sample of 45 forecasts to assess the performance of the scheme in operation.

We first assess the skill of the seasonal forecasts. Due to the limited number of years of operation, we combine seasonal forecasts issued in three different months to evaluate the skill for each season. For instance, the ‘winter’ skill is evaluated using the seasonal forecasts for the three periods NDJ, DJF, and JFM. By comparison, only the DJF period was evaluated in the hindcasts. This increases the sample size while mitigating against the variations over the year identified in Fig. 4(b).

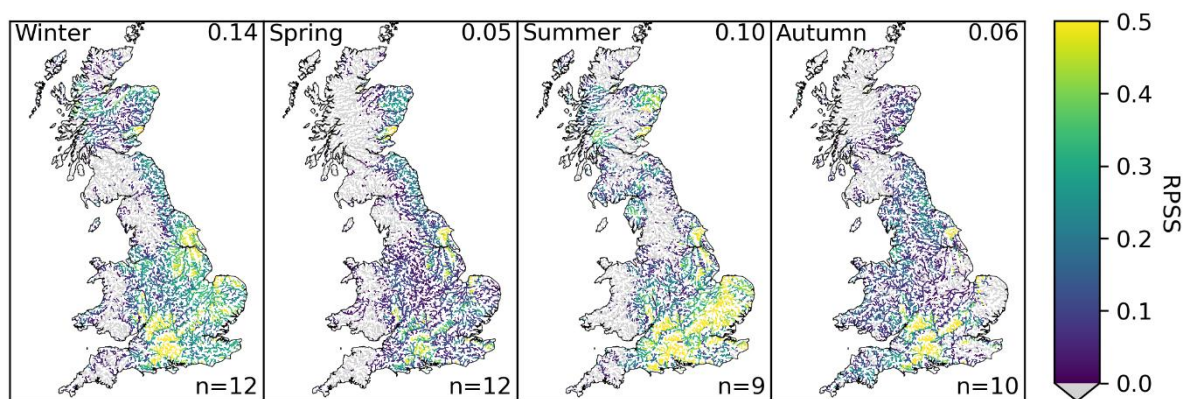


Figure S2: Ranked Probability Skill Scores (RPSS) showing the value added by HWA-driven river flow forecasts over climatology during the commissioning and operational periods. Coloured river pixels show that the river flow forecasts outperform climatology ($RPSS > 0$). The GB-mean RPSS is shown in the top-right of each panel, and the number of seasonal forecasts is in the bottom-right.

The Ranked Probability Skill Scores (RPSS) of the operational seasonal forecasts is shown in Figure S2. The equivalent plots for the hindcasts are in the top row of Fig. 4b. These broadly follow the spatial patterns identified for the hindcasts, with the greatest skill seen in the south-east of England. In winter, the skill in the north-west of Scotland can be attributed to the HWA rainfall forecasts as previously shown in Fig. 7a. Some small differences can be identified. For instance, forecasts in Wales in winter and spring are not skilful in the operational scheme but were marginally skilful ($0 < RPSS < 0.1$) in the hindcasts. Given the small size of the operational sample such small variations in RPSS are not thought to be significant.

The GB-mean skill scores are very similar to those in the hindcasts (Table S1). The notable exception is in the summer, where the operational forecasts have performed substantially better than the hindcasts in the south-east. We attribute this to the low summer rainfall experienced in these parts recently, during which hydrological memory was even more influential than usual, leading to more predictable river flows. In spring, more river cells have skilful forecasts, but the spatially averaged skill has scarcely changed between hindcasts and forecasts.

Table S1: Comparison between the Ranked Probability Skill Scores (RPSS) for the hindcasts (1993-2017) and operational (2022-2026) seasonal forecasting systems.

Season	GB-mean RPSS		% of river cells with RPSS>0	
	Hindcasts	Operational	Hindcasts	Operational
DJF/Winter	0.12	0.14	68%	73%
MAM/Spring	0.04	0.05	45%	63%
JJA/Summer	-0.06	0.10	20%	64%
SON/Autumn	0.04	0.06	60%	67%

It is illustrative to compare the forecasts to simulated observed flows at an individual site. Figure S3 shows an example for Thames at Kingston (National River Flow Archive station 39001), a site with a long history in UK hydrology. We see that forecast variability changes throughout the year. In the winter, forecasts are less confident (indicated by the larger grey envelopes) because of the greater variability in flows. In the summer, flows are less variable, with groundwater contributions dominating, and thus the forecasts are more confident (indicated by narrower grey envelopes).

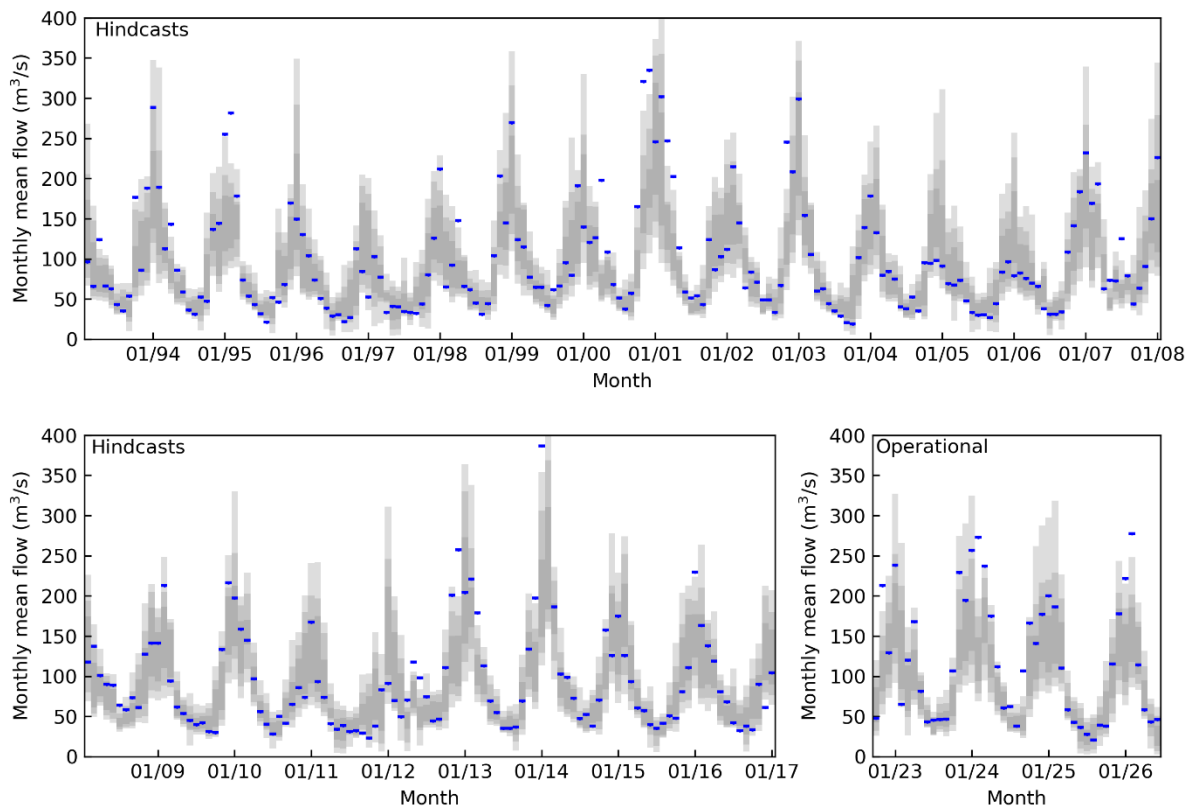


Figure S3: Monthly hindcasts (top and bottom left panels) and operational forecasts (bottom right panel) for Thames at Kingston. Blue lines indicate the simulated observed flows estimated with the Grid-to-Grid hydrological model, while the grey envelopes encompass the 50%, 75% and 90% confidence interval of each forecast.

We also observe that simulated observed flows (blue lines) usually fall within the forecast envelope. Table S2 quantifies the proportion of months in which flows fall within each forecast confidence interval. We see little difference between the hindcasts and forecasts at this site.

Table S2: Proportion of monthly forecasts for which the simulated observed flows are within given confidence intervals for the Thames at Kingston.

Confidence interval	Hindcasts	Operational
50%	51%	53%
75%	79%	69%
90%	91%	86%

Although only a limited number of recent forecasts are available, these results give us confidence that the conclusions drawn about the skill of the forecasting system based on the 1993-2017 hindcast sample remain robust into the early and mid-2020s. Continuous performance assessment is part of the production process for the operational forecasts, and so over the coming years this will continue to be assessed by the forecasting team.