Received: 12 January 2021 | Revised: 20 February 2022 | Accepted: 22 February 2022 | Published on: 29 March 2022

DOI: 10.1002/qj.4259

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Influence of trends on subseasonal temperature prediction skill

C. Ole Wulff[1,2] | **Frédéric Vitart**[2,3] | **Daniela I. V. Domeisen**[1,4]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

[2]NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway

[3]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

[4]University of Lausanne, Lausanne, Switzerland

Correspondence
C. Ole Wulff, NORCE Norwegian Research Centre AS, Postboks 22 Nygårdstangen, NO-5838 Bergen, Norway.
Email: owul@norceresearch.no

Funding information
Swiss National Science Foundation, Grant/Award Number: PP00P2_170523

Abstract

Subseasonal-to-seasonal (S2S) predictions have a wide range of applications. Improving forecasts on this time-scale has therefore become a major effort. To evaluate their performance, these forecasts are routinely compared to a reference that forecasts the climatological distribution at any given time. This distribution is commonly assumed to be stationary over the verification period on time-scales longer than the seasonal cycle. However, there are prominent deviations from this assumption, especially considering trends associated with climate change. By employing synthetic forecast-verification pairs, we show that estimates of the probabilistic skill of both continuous and categorical forecasts increase as a function of the variance explained by the trend over the verification period, even when there are errors in the trends simulated by the forecasts. We also show this skill enhancement due to the trend in the ECMWF extended-range ensemble prediction system. We demonstrate that the effects on the skill in an operational forecast setting are currently strongest in the Tropics. Our results show that care needs to be taken when evaluating forecasts that are subject to non-stationarity on time-scales much longer than the forecast verification window. This is especially important for determining the skill of categorical forecasts, where assumptions on the stationarity of the climatology enter both in the reference forecast and in the determination of the category thresholds. The results presented in this study are not exclusive to the S2S time-scale but have wider implications for forecast verification on seasonal to decadal time-scales, where the existence of trends can further impact forecast skill.

KEYWORDS

probabilistic skill, subseasonal forecasts, trend, verification

1 | INTRODUCTION

Stakeholders face decisions related to weather and climate risks on a continuum of time-scales from minutes into the future to multiple decades or even centuries. In recent years, increasing efforts have been made to move towards 'seamless' prediction (Hoskins, 2013) to bridge the gap between classical weather forecasting for

lead times of days and climate projections for the next century. The potential for successful predictions on all time-scales stems from a multitude of large-scale phenomena in the atmosphere and the ocean that evolve in different parts of this temporal spectrum. One case in point is the El Niño–Southern Oscillation (ENSO) – a coupled ocean–atmosphere phenomenon with a variable frequency between 2 and 7 years that gives rise to seasonal prediction skill not only in its region of occurrence but also in the midlatitudes of the globe through atmospheric teleconnections (Shukla *et al.*, 2000). Other sources of predictability can be found on time-scales from weeks to months, the so-called subseasonal-to-seasonal (S2S) time-scales (Vitart *et al.*, 2017 give a general overview), for instance the Madden–Julian Oscillation (MJO; e.g., Lee *et al.*, 2019), stratospheric processes (e.g., Domeisen *et al.*, 2020), land surface processes (e.g., Koster *et al.*, 2011), and sea ice variability (e.g., Jung *et al.*, 2014). The S2S time-scale has furthermore been shown to be particularly relevant for decision makers (White *et al.*, 2017; Robertson *et al.*, 2020). We will focus on forecasts on these time-scales in this study.

Despite the potential for predictability, S2S forecasts generally exhibit significantly lower skill than forecasts on weather time-scales. It is thus crucial to use ensemble systems in order to sample the space of possible outcomes given the uncertainty in the estimate of the initial state (Leutbecher and Palmer, 2008). Despite the low deterministic skill on S2S time-scales, these probabilistic forecasts can be useful in making decisions that depend on the weather evolution, given that they quantify the uncertainty in the outcomes correctly. Probabilistic subseasonal forecasts have, in fact, been shown to be skilful in a variety of settings (e.g., Alvarez *et al.*, 2020; Materia *et al.*, 2020; Robertson *et al.*, 2020).

In order to increase users' confidence in S2S forecasts, it is imperative to show under which circumstances these can provide useful information. In the scientific literature, skill scores are often used as measures of the quality of a forecast. A skill score relates some measure of accuracy (e.g., the mean error with respect to observations in a location) of the forecast under consideration to the same measure of a reference forecast (Wilks, 2019). It thus expresses the improvement of a forecast over this reference. In subseasonal forecasting, the most typical reference is a climatological forecast, which represents the distribution of all possible outcomes of a physical quantity given the current climate. However, the true climatological distribution is not known and needs to be estimated from the available data. Since there is no unique way of estimating the climatology and different choices might be appropriate in different settings, skill scores defined with respect to climatology can vary substantially depending on the assumptions made about the climate. In a conceptual way this was shown previously by Hamill and Juras (2006) using synthetic forecast–observation pairs. They introduced two hypothetical islands with different climatological event frequencies to illustrate how a forecasting system that always issues the climatological frequency for each respective island (and thus has no actual skill) can appear to have skill when the aggregated climatology for both islands is used as reference. While the solution to the artificial enhancement of the skill in this hypothetical example is straightforward, in an operational setting it might not be trivial to estimate the climatology due to the available sample size. Manrique-Suñén *et al.*, (2020) used the operational extended-range predictions from the European Centre for Medium-Range Weather Forecasting (ECMWF) and found that there can be substantial differences in the estimates of the skill of subseasonal forecasts depending on the chosen method to compute the climatology from the limited hindcast period.

In a general sense, the aforementioned results illustrate the importance of properly accounting for the non-stationary components in the climatological distribution when assessing the skill of a forecast with respect to climatology (also DelSole and Tippett, 2018). While the seasonal cycle is often accounted for in the reference climatology for the evaluation of forecasts, non-stationary components that act on longer time-scales are commonly neglected. One prominent example of a non-stationary component in temperature is global warming (IPCC, 2013). For dynamical seasonal forecasting systems to produce realistic warming, the greenhouse gas forcing giving rise to it needs to be accounted for in the boundary conditions (Doblas-Reyes *et al.*, 2006; Liniger *et al.*, 2007; Boer, 2009). In statistical predictions, trends have been shown to be among the most important predictors for North American monthly to seasonal temperatures (Peng *et al.*, 2012; Johnson *et al.*, 2014). While the magnitude of global warming shows large regional variability, it is manifest in temperature time series throughout the globe. In many places, a shift in the mean temperature can be detected even when considering only the recent past, e.g. the last 30 years, which is a common period for defining a climatology. Based on the arguments above, this non-stationary component of the climatology on time-scales longer than the seasonal cycle, has the potential to affect the estimates of subseasonal forecast skill.

Our study aims to characterise and quantify the effect of a trend in the climatological reference period on the probabilistic skill of subseasonal forecasts. We test if there is an enhancement of skill and assess how strongly the magnitude of the improvement depends on the amount of variance of the respective time series that the trend accounts for. We introduce the forecast and verification

data and the processing methods in Section 2 along with the probabilistic scores for evaluating their performance. In Section 3, we quantify the effect of a trend on different probabilistic skill scores in a set of synthetic forecast–verification pairs to separate the stationary from the non-stationary component. Here, we also assess what happens to the skill scores when there is a trend in the verification data but the forecasts fail to reproduce it correctly. We then compare the behaviour of an operational prediction system (the ECMWF's extended-range ensemble prediction system) to our synthetic model in Section 4 and show in which geographical areas the subseasonal forecast skill may be most strongly affected by the presence of a trend in the mean of the climatological distribution. We review the results in the context of previous literature and discuss limitations of the synthetic model in Section 5. Finally, we present the conclusions of our study in Section 6.

2 | DATA AND METHODS

2.1 | Forecast data and verification

For assessing the trend effect in an operational ensemble prediction system, we make use of the extended-range forecasting system of the ECMWF. This extended-range forecast ensemble is operationally produced with the most recent version of the ECMWF Integrated Forecasting System (IFS) by extending the weather forecast runs twice a week out to 46 days (instead of 15). This is done by re-starting the forecast runs on day 14 at a reduced horizontal resolution (Tco319 instead of Tco639), but note that this first day of the re-start is only used as spin-up. The horizontal resolution corresponds to approximately 16 km up to day 15 and 32 km after this, and the model has 91 vertical levels. The ensemble in the operational setting consists of 51 members. For each forecast, hindcast ensembles with 11 members are produced by initialising the same model version from re-analysis (ERA-Interim up to and including IFS cycle 45R1, ERA5 from cycle 46R1) on the same calendar day for the previous 20 years. The atmospheric component of the IFS is coupled to an ocean and an interactive sea ice model and uses the HTESSEL land surface scheme. The IFS uses a boundary forcing with varying greenhouse gas (GHG) concentrations following the Coupled Model Intercomparison Project (CMIP3) A1B scenario (Meehl *et al.*, 2007). Since the GHG forcing thus does not only enter through the initial conditions, we expect the model to produce realistic trends at all potential lead times.

To cover a period that is sufficiently long to consider an effect of trends on the forecasts, we retrieved 20 years of hindcast data through the S2S database (Vitart *et al.*, 2017). We downloaded daily mean 2 m temperatures (T_{2m}) from the ECMWF's extended range ensemble forecasts initialised between 1 January 2018 and 31 December 2018 (twice-weekly initialisation, giving 105 forecasts) as well as the corresponding hindcasts (same initialisation days within the year for the period 1998–2017). For each initialisation, 20 years of hindcasts are produced, yielding a sample of 2,100 hindcast–observation pairs for each lead time. In addition, we extended the number of samples in the forecast period by including all forecasts initialised between 1 January 2018 and 1 January 2021 resulting in 315 initialisations.

Note that, in order to increase the sample size for the forecasts and hindcasts, we use varying model versions in our analysis. In particular, for this period of hindcasts and forecasts, the ECMWF changed from cycle 43R3 to 45R1 to 46R1 to 47R1 of the IFS (https://confluence.ecmwf.int/display/S2S/ECMWF+Model#app-switcher; accessed 14 March 2022) and thus the model data considered here were generated with four versions of the IFS. There are some important differences in the IFS between versions, which impact the forecast and hindcast skill to some degree, but the effect of the changes in model version used here has been mainly visible on the shorter lead times of the forecasts (e.g., Vitart *et al.*, 2019).

As verification data for the hindcasts and forecasts, we use daily mean 2 m temperatures from ERA5 (Hersbach *et al.*, 2020). The data were downloaded globally at $1° \times 1°$ resolution for the period 1 January 1997 – 28 February 2021.

2.2 | Estimating the seasonal cycle

For the remainder of this study, it is useful to consider standardised temperature anomalies because it allows us to easily identify the amount of variance that different components of the time series account for in the hindcast period. By accounting for the model's own climatological mean and standard deviation, it additionally ensures a certain degree of calibration of the forecasts, but note that for a proper calibration, a more sophisticated approach would be required. For our simple calibration, we compute the seasonal temperature cycle from the hindcast period (1998–2017) only.

To transform to standardised anomalies, we need to estimate the climatological seasonal cycle of the mean and standard deviation. In a first step, we estimate the mean temperature $\overline{T}_{i,l}$ on each of the 105 initialisation dates i per year and for each lead time l:

$$\overline{T}_{i,l} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{M}\sum_{m=1}^{M} T_{m,n,i,l}\ , \tag{1}$$

where $T_{m,n,i,l}$ refers to the 7-day mean, detrended temperature, index $m \in \{1, 2, ..., M\}$ indicates the ensemble member ($M = 51$ for the forecasts and $M = 11$ for the hindcasts), and index $n \in \{1, 2, ..., N\}$ denotes the year in the hindcast period ($N = 20$). In order to obtain a smooth seasonal cycle for the climatological mean, we then retrieve the lowest four harmonics of $\overline{T}_{i,l}$ for each l individually, retaining only variations with periods greater than approximately 90 days. We refer to the climatological seasonal cycle of the mean, reconstructed from these first four harmonics as $\widetilde{\overline{T}}_{i,l}$. Temperature anomalies are then defined relative to the seasonal cycle as

$$T'_{m,n,i,l} = T_{m,n,i,l} - \widetilde{\overline{T}}_{i,l}. \tag{2}$$

In the next step, we estimate the seasonal cycle of the standard deviation. We first compute the standard deviation of the anomalies for each initialisation date i and lead time l as

$$s_{T'_{i,l}} = \sqrt{\frac{1}{(MN)-1}\sum_{n=1}^{N}\sum_{m=1}^{M}(T'_{m,n,i,l})^2}. \tag{3}$$

As above, we next estimate the first four harmonics of $s_{T'_{i,l}}$ for each l individually and reconstruct a smooth seasonal cycle of the standard deviation, which we refer to as $\widetilde{s}_{T'_{i,l}}$. Dimensionless, standardised temperature anomalies $T^*_{m,n,i,l}$ are then computed as

$$T^*_{m,n,i,l} = \frac{T'_{m,n,i,l}}{\widetilde{s}_{T'_{i,l}}}. \tag{4}$$

2.3 | Processing of hindcasts and verification

We first average the hindcast, forecast and ERA5 temperatures to 7-day averages. In those cases where the data are detrended, the detrending step precedes the standardisation. For the detrending, we compute an annual mean linear trend over the hindcast period (1998–2017), which is subtracted from the absolute temperatures. We then transform the detrended fields into dimensionless standardised anomalies according to Equation (4) to minimise contributions from the seasonal cycle to the hindcast skill.

Note that, in the case of the forecasts, the seasonal cycle (mean and standard deviation) as well as the trend are a function of the lead time to account for possible drifts in the model climatology.

For transforming the ERA5 temperatures to standardised anomalies, we follow the same approach as outlined above for the hindcasts. This means we take ERA5 temperatures for the same 20 years as the hindcasts (1 January 1998 to 31 December 2017) averaged over 7-day periods. The seasonal cycles in mean and standard deviation are then computed as in Section 2.2 (with $M = 1$ and without the need to account for a lead time dimension l) but using all 365 days of the year.

Due to the filtering in the above-described estimation of the seasonally varying climatology, the mean and variance of the standardised anomalies over the hindcast period deviate from zero and one, respectively. Since it is mainly their variance that deviates from unity, in a last step, we standardise the time series again by dividing by the empirical standard deviation of the standardised anomalies over the entire hindcast period. This ensures unit variance over the hindcast period for both the verification and hindcasts at all lead times.

2.4 | Forecast verification

2.4.1 | Scoring

There exists a multitude of scores that can be used to assess the performance of a forecast (e.g., Jolliffe and Stephenson, 2012). Since there is no single measure that captures all aspects of the performance, usually multiple measures are applied and the choice of the scores often depends on the specific application. Here, we use a fairly general score for the evaluation of forecasts of continuous variables, the continuous ranked probability score (CRPS; Wilks (2019)), which summarises multiple attributes of a forecasting system. For a forecast of temperatures y at a single instance when a temperature o was observed, the CRPS is given by:

$$\mathrm{CRPS} = \int_{-\infty}^{\infty}[F(y) - F_o(y)]^2 \mathrm{d}y, \tag{5}$$

where $F(y)$ is the cumulative distribution function (CDF) of the temperatures in the forecast ensemble and $F_o(y)$ is the CDF of the observations given by:

$$F_o(y) = \begin{cases} 0, & y < o, \\ 1, & y \geq o, \end{cases} \tag{6}$$

which describes a step function with a jump from 0 to 1 at temperature $y = o$.

The CRPS describes the error of a probabilistic forecast by the integrated squared distance between the forecast and the observed CDFs. The CRPS is negatively oriented,

meaning the smaller the score, the better the forecast. In all cases considered, we average the CRPS over the number of all forecast–verification pairs.

In case $F(y)$ is known, the CRPS can be computed directly from Equation (5). This will be the case for our climatology forecasts that predict a standard normal distribution at any time. Because the actual forecasts (both in Sections 3 and 4) have a finite number of ensemble members, we estimate their score using the kernel representation of the (adjusted) CRPS as presented in Leutbecher (2019). This adjusted version of the score accounts for the effects of having a finite number of ensemble members and is thus a more fair comparison to the reference score.

We also employ a categorical score to evaluate the performance of the forecasts. For this, we use the RPS as defined by Wilks (2019):

$$\mathrm{RPS} = \sum_{k=1}^{K} (Y_k - O_k)^2, \tag{7}$$

where K is the number of categories that verifications and forecasts are sorted into. The observation O_k attains a value of 0 or 1 if the verification at the considered time step lies outside of or inside category k, respectively. The forecast probability Y_k can have values between 0 and 1 and indicates what fraction of ensemble members were in category k. Choosing $K = 3$ equiprobable categories makes the RPS the natural choice for evaluating tercile forecasts. This requires the tercile thresholds to be defined. In the case of the synthetic model (Section 3.1.1) we know the exact values for these thresholds at any time step since we prescribe the distribution from which the forecasts are drawn. In reality however, the thresholds have to be estimated from the climatology. We discuss the issue of estimating the thresholds from the hindcast period further in Section 3.2.

Like the CRPS, the RPS shows some dependence on the ensemble size (e.g., Richardson, 2001; Müller *et al.*, 2005) that reflects the 'intrinsic unreliability' of an ensemble with a finite number of members (Weigel *et al.*, 2007). Following Ferro *et al.*, (2008), this effect can be approximately accounted for by scaling the RPS of the forecasts and hindcasts with a factor $D = M/M + 1$, which allows for a fair comparison with the RPS of the reference forecast.

2.4.2 | Skill

In the following sections, we further evaluate the skill of the forecast by considering the relative improvement in its score S over the score $\mathrm{S_{ref}}$ of a reference forecast. For this, we define a skill score SKS, which can vary between $-\infty$ and 1, where 0 indicates no improvement over the reference and 1 means that the score S attains its optimal value ($\mathrm{S_{opt}} = 0$, for all scores considered here). The skill score SKS is thus given by:

$$\mathrm{SKS} = \frac{\overline{\mathrm{S}} - \overline{\mathrm{S_{ref}}}}{\overline{\mathrm{S_{opt}}} - \overline{\mathrm{S_{ref}}}} = 1 - \frac{\overline{\mathrm{S}}}{\overline{\mathrm{S_{ref}}}}. \tag{8}$$

Here, overbars denote the average over all forecast–verification pairs. Since we only consider averages over the entire sample (either hindcasts or forecasts), the overbars are dropped in the following. In our case, S is either the CRPS or the RPS and SKS is the CRPSS or RPSS, respectively.

As can be seen from Equation (8), the skill score depends on S_{ref}, the score of the reference forecast, and is thus sensitive to the definition of the reference forecast itself. A common choice in the verification of subseasonal forecasts is a climatological reference forecast. In the following sections, we will show in detail how a non-stationarity in the climatology in the form of a linear trend can affect the score of the reference and thus the skill scores of subseasonal forecasts.

3 | DEPENDENCE OF THE SKILL ON UNDERLYING TRENDS: SYNTHETIC ENSEMBLE SYSTEM

In the following, we consider a linear trend in a verification time series to test how the probabilistic skill of a simple hypothetical ensemble prediction system changes as a function of the magnitude of the trend. We define a set of synthetic forecast–verification pairs to be able to cleanly separate the stationary random and predictable parts of the time series from the non-stationary component (the linear trend). Since any real forecasting system is also subject to errors, we relax the assumption of a perfect simulation of the trend in the forecast and employ the toy forecast to assess how such an imperfect estimation of the trend can further affect the skill. This assessment provides a benchmark for the potential magnitude of the effect that a trend in the forecast period can have on the skill of the forecast ensemble depending on the trend magnitude and the mis-estimation.

3.1 | Set-up of the artificial forecast–verification pairs

In the next sections, we describe the synthetic ensemble forecast–verification pairs that we generate in order to answer the questions posed above. We follow the approach of Weigel *et al.*, (2008) who used a similar toy forecast model (without a trend but with a parameter controlling

the forecast dispersion) to study why multi-model ensembles can outperform the single best models. We use the notation $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2.

3.1.1 | Verification time series

We first generate an artificial verification time series v_t consisting of a predictable signal ϕ_t, an unpredictable noise component ϵ_t and a linear trend Δ_t:

$$v_t = \Delta_t + \phi_t + \epsilon_t, \tag{9}$$

with index $t \in \{0, 1, ..., L_{\text{hc}}, L_{\text{hc}} + 1, ...L\}$ indicating the time step and L the length of the entire time series. The first interval $[0, L_{\text{hc}} - 1]$ has length L_{hc} and is referred to as the hindcast period. The interval $[L_{\text{hc}}, L]$ of length $L_{\text{fc}} = L - L_{\text{hc}} + 1$ is the forecast period.

Δ_t is a linear trend defined as

$$\Delta_t = \gamma \left(t - \frac{L_{\text{hc}}}{2} \right), \tag{10}$$

with γ being the slope of the trend. By definition, the trend line has mean $\mu_\Delta = 0$ in the hindcast period $[0, L_{\text{hc}} - 1]$. The variance contribution of the trend line in the hindcast period is a function of the slope γ and the length of the hindcast period L_{hc} only:

$$\sigma_\Delta^2 = \frac{(\gamma L_{\text{hc}})^2}{12}. \tag{11}$$

We let both the signal ϕ and the noise ϵ be white noise with values for each t drawn from $\mathcal{N}(0, \sigma_\phi^2)$ and $\mathcal{N}(0, \sigma_\epsilon^2)$, respectively. The combination of signal ϕ and noise ϵ represents the detrended part of the verification, which has variance $\sigma_x^2 = \sigma_\phi^2 + \sigma_\epsilon^2$. We want to ensure unit variance of v over the hindcast period, which requires

$$\sigma_x^2 = 1 - \sigma_\Delta^2. \tag{12}$$

We further choose

$$\sigma_\phi^2 = \alpha^2 \sigma_x^2, \tag{13}$$

where $0 \leq \alpha^2 < 1$ defines the fraction of the detrended variance contained in the signal (Section 3.1.2 has a further discussion of the meaning of α).

3.1.2 | Ensemble prediction system

We generate a prediction ensemble with M members at a time step t in a similar fashion to Weigel *et al.*, (2008). In the following, we refer to the predictions for the hindcast period ($[0, L_{\text{hc}} - 1]$) as the hindcasts and the predictions for the forecast period ($[L_{\text{hc}}, L]$) as the forecasts. We let each ensemble member f_m ($m \in [1, 2, \ldots, M]$) predict the signal ϕ, a trend component τ and Gaussian noise $\varepsilon_m \sim \mathcal{N}(0, \sigma_\varepsilon^2)$:

$$\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_M \end{pmatrix}_t = \phi_t + \tau_t + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{pmatrix}_t. \tag{14}$$

The trend τ_t of the prediction is defined as the observed trend Δ_t multiplied by a factor p, which represents the mis-estimation of the trend in the prediction system:

$$\tau_t = p \Delta_t. \tag{15}$$

To achieve calibration of the predictions to the verification, we set the variance of the hindcasts to $\sigma_{f_m}^2 = 1$. Since the variance of each ensemble member is given by $\sigma_{f_m}^2 = \sigma_\phi^2 + p^2 \sigma_\Delta^2 + \sigma_\varepsilon^2$, conditions (11) and (13), together with $\sigma_{f_m}^2 = 1$, imply

$$\sigma_\varepsilon^2 = 1 - \left(p^2 \frac{(\gamma L_{\text{hc}})^2}{12} + \alpha^2 \sigma_x^2 \right). \tag{16}$$

In this way, we ensure that hindcasts have zero mean and unit variance, as is the case for the verification in the hindcast period. Note that, due to setting the variance of hindcasts and verification to 1, the variances of the different components in the verification and predictions are equivalent to the fraction of explained variance in the hindcast period of the respective time series. For instance, σ_Δ^2 describes the fraction of variance of the verification accounted for by the trend during the hindcast period, while $\alpha^2 \sigma_x^2$ is the fraction of variance explained by the predicted, detrended part.

At this point, we want to briefly discuss the meaning of the parameter α. Since the predicted part of the variance of the verification and the hindcast ensemble in the absence of a trend ($\sigma_\Delta^2 = 0$) is α^2, α can be understood as the correlation between the detrended verification and the ensemble mean hindcast. By setting α, we thus set the theoretical skill of our synthetic prediction system at predicting variability on the time-scale of interest. We will refer to α as the detrended correlation skill of the system. Note that, once a trend is introduced to the model, the actual correlation between the verification and the hindcast's ensemble mean increases. Since a trend constitutes a perfectly predictable component in the time series, it will impact other measures of forecast skill (Section 2.4) as well. Our simple model allows us to separate the skill arising due to the changing climatology from the skill at forecasting

variability on the time-scale of interest. In the operational forecasts considered in Section 4, these time-scales are subseasonal but note that the definition of the synthetic model is general enough for the results to be applicable to other time-scales.

In a last step, we have to choose the parameters of our model. Our aim is to mimic the set-up of the actual prediction system. At the same time, since the uncertainty of the scores is smaller for a larger sample size, we aim for a larger sample than in the operational system to obtain a more robust estimate. Since we suspect the trend effect to partially depend on the length of the forecast period relative to the length of the hindcast period, we keep this ratio the same by letting our synthetic prediction system have the same number of forecast and hindcast years as the operational system (3 and 20, respectively), but with more initialisations in each year. Thus, we set $L_{\mathrm{fc}} = 1{,}050$ and $L_{\mathrm{hc}} = (20/3)L_{\mathrm{fc}} = 7{,}000$. Consistent with the operational system, we only use $M_{\mathrm{hc}} = 11$ ensemble members for the ensemble hindcasts but $M_{\mathrm{fc}} = 51$ members for the forecasts. The remaining parameters, γ, α and p are varied. Setting L_{hc} and fixing the variance of the verification to 1 in the hindcast period sets a limit to the possible values of the slope of the trend γ, which is readily understood when we consider σ_{Δ}^2 as the fraction of variance explained by the trend; this value cannot exceed 1. Thus, we obtain $0 \leq \gamma < \sqrt{12}/L_{\mathrm{hc}}$. Since α represents the correlation between verification and prediction, we let the prediction system vary from having no skill at all to having nearly perfect correlation with the verification, that is, $0 \leq \alpha < 1$. Finally, the mis-estimation factor p of the trend can be varied. The unit variance of the hindcasts allows us to vary it within the limits $0 \leq p < {\sigma_{\Delta}}^{-1}\sqrt{1-\alpha^2\sigma_x^2}$.

Note that, when using the model, for every combination of parameters, we use the same verification and forecast time series. In practice, this means that we draw L values for each component in Equations 9 and 15 only once and scale them according to the variance ratios given above, which depend on the chosen parameters.

3.2 | Effects of varying trend and detrended skill on probabilistic skill scores

3.2.1 | An illustrative example

To illustrate the expected effect of a trend in the verification, consider the synthetic verification time series shown in Figure 1a, which is an extension of figure 10.2 in Livezey (1999). In this case, the trend explains 6% of the variance in the hindcast period. At each time step t, the verification is a single draw from a normal distribution $\mathcal{N}(\Delta_t, \sigma_x^2)$. The tercile thresholds of this distribution as functions of t are shown by the green dashed lines. However, in a real forecast situation, the tercile thresholds have to be estimated from the hindcasts. It is common to use all time steps in the hindcast period for estimating the climatological distribution. Due to the way we defined the verification, this distribution has mean zero and unit standard deviation. The tercile thresholds of a corresponding standard normal distribution (± 0.431) are shown by the black solid lines. Figure 1b illustrates what happens when these thresholds are used to define the tercile categories for the forecasts. For this, we split the time series by years (one year consisting of 350 time steps). Only in the centre of the hindcast period, approximately 1/3 of values fall into each category. Towards the beginning, 50% of values end up in the lower tercile, while towards the end of the hindcast period, almost half of the values are sorted into the upper tercile. Since it lies after the hindcast period, this effect is even stronger in the forecast period where more than 50% of values are in the upper tercile even though the trend only explains 6% of the variance. If we now imagine a forecast which is able to reproduce this trend correctly and has realistic dispersion (as our synthetic predictions) but no skill at predicting any other part of the detrended variability in the verification, it similarly sorts more than a third of the forecast values into the upper tercile. When evaluated with a categorical score such as the RPS, the score for the hindcasts will inevitably be better than if we had used the changing percentiles (green dashed lines in Figure 1a) of the verification. Note especially that the score of the forecasts will be better than the score of the hindcasts despite the forecasts not having any more actual skill. Finally, the RPS also improves as the forecast period over which the skill is evaluated is extended. This illustrates the problem we face when evaluating categorical forecasts over a period that is subject to a trend using empirical category thresholds estimated from the hindcast period.

3.2.2 | Quantitative analysis

To assess the effect of any underlying trend on different probabilistic scores of the toy model and their skill estimates for different levels of correlation skill α, we now vary the prescribed slope γ of the trend. For the following results, we assume that the prediction system simulates the observed trend Δ_t perfectly, that is, $p = 1$ and thus $\tau_t = \Delta_t$. We then compute the average CRPS and RPS over all time steps t as described in Section 2.4 in both the hindcast and forecast periods. Note that for the computation of the RPS (Equation (7)), we need to define the tercile thresholds. To highlight the effect described above, we show both the RPS estimated with a constant tercile threshold over the entire period (as would be estimated under the

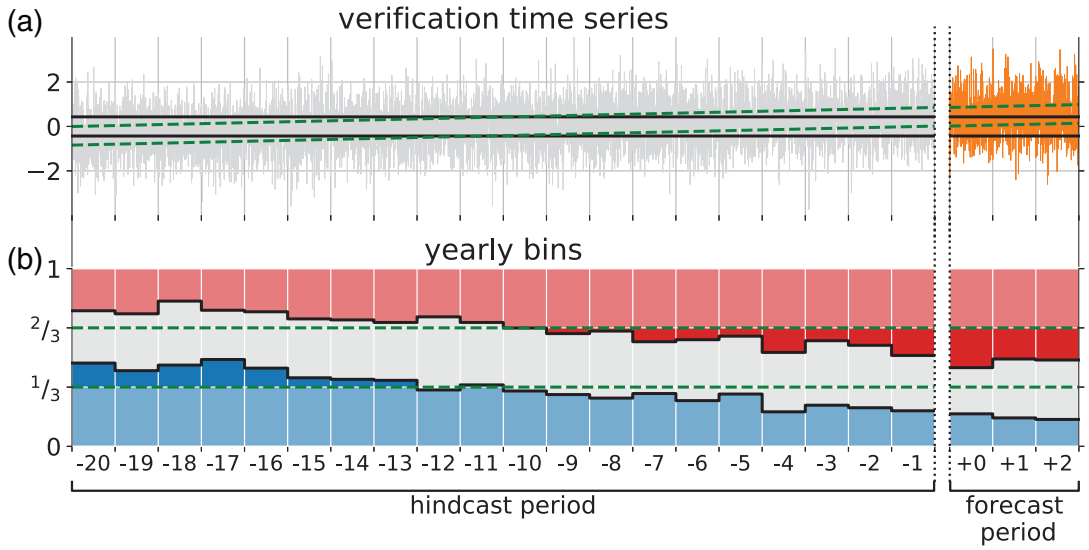

FIGURE 1 Illustration of trend effect on a categorical forecast. (a) shows a synthetic verification time series generated as described in Section 3.1.1. The grey line indicates the hindcast period and the orange part the forecast period (separated by dotted lines). The black solid lines show the 33.3 and 66.7 percentiles of the distribution of a standard normal distribution. Green dashed lines show the same percentiles but accounting for the underlying trend which here explains 6% of the variance in the hindcast period. (b) shows the fraction of time steps in each year (consisting of 350 time steps) in which the verification falls into the lower (blue), middle (grey) and upper (red) tercile of a standard normal distribution based on the percentile thresholds shown by the black solid lines in (a). Green dashed lines indicate the theoretical tercile bins for tercile thresholds that account for the trend, corresponding to the green dashed lines in (a) [Colour figure can be viewed at wileyonlinelibrary.com]

assumption of a stationary climatology) and the RPS estimated with the changing tercile thresholds that increase monotonically over time, following the trend. We furthermore compute the skill scores (Equation (8)), namely the CRPSS and the RPSS. This requires us to define a reference, which we choose to be a climatological forecast. This forecast predicts the climatological distribution for every time step t, which is assumed to be Gaussian in all of our analyses. Under the assumption of a stationary climatology, the predicted distribution is the same for every t and given by $\mathcal{N}(0,1)$ since we consider standardised anomalies. If we account for the changing climatology however, the climatological forecast for every t is given by $\mathcal{N}(\Delta_t, \sigma_x^2)$. In the following, we compare the skill computed with respect to either of these references.

We first consider the effects of defining the climatology for the CRPS and CRPSS (Figure 2a–d). The CRPS itself does not depend on the definition of the climatology, which is manifest in the fact that the black contour lines are the same for Figure 2a–d. Comparing the black contours with the blue contour lines, which show the prescribed detrended skill α of our system, it also becomes clear that the CRPS does not follow α but instead decreases with stronger trends (i.e., larger amounts of variance explained by the trend). This is unsurprising and similarly, if we were to compute a correlation between raw forecasts and verification, we would observe the same effect, namely the correlation increasing with a stronger trend. We next compute the skill score, namely the CRPSS, which is shown by the shading in Figure 2a–d. In Figure 2a, c, the CRPSS is computed with respect to a reference forecast that assumes a stationary climatology. Clearly, the skill score in these panels is not aligned with the detrended skill (blue contours) either. For the hindcasts (Figure 2c), the CRPSS follows the decrease in the CRPS exactly. For the forecasts (Figure 2a), it increases even more strongly as a function of the trend. The reason is that the stationary climatology is a worse model in the forecast period than in the hindcast period. While verification anomalies in the hindcast period all lie well within the stationary climatological distribution ($\mathcal{N}(0,1)$), verification anomalies in the forecast period will regularly lie far in the upper tail of the assumed climatological distribution when there is a trend. These 'outliers' will occur more often if the trend is stronger, making the stationary climatology a much less competitive model during the forecast period. This again shows the issue of a trend in the verification period: the synthetic prediction model as we define it does not make better forecasts for any period, and its skill is stationary. Yet, when the skill is evaluated under the assumption of a stationary climatology, it will be computed to be higher during periods that lie outside of the hindcast period over which the climatological distribution is estimated. To compensate for this effect, it is possible to use a changing

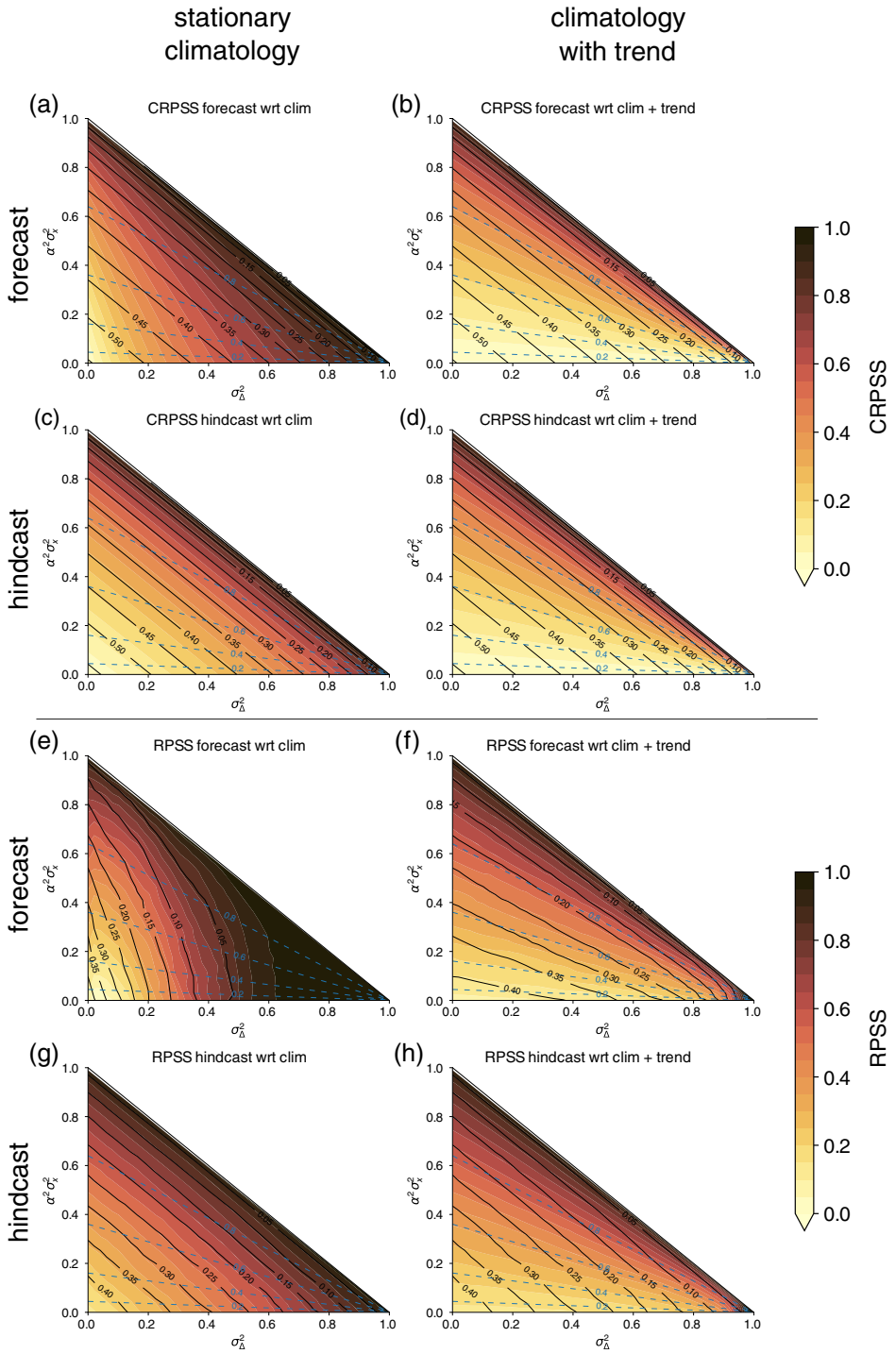

FIGURE 2 Dependence of skill scores on the choice of the reference climatology in the synthetic model. In all panels, the skill score (CRPSS in a–d, RPSS in e–h) is depicted by the shading. The reference score in (a, c, e, g) is computed from a climatological forecast assuming a long-term stationary climatology. In (b, d, f, h), the long-term trend is subtracted from the climatology. The respective score (CRPS in a–d, RPS in e–h) of the forecasts is shown by the black contours. In (a, b, e, f) [c, d, g, h] the synthetic forecasts are verified over the forecast [hindcast] period. Blue dashed lines indicate the prescribed level of skill α [Colour figure can be viewed at wileyonlinelibrary.com]

climatology as a reference forecast instead. To illustrate that this compensation works, the shading in Figure 2b, c shows the CRPSS with respect to a reference forecast that accounts for the trend in the climatology. The thus defined CRPSS follows the detrended skill of the system (compare shading to blue contours) exactly and thus is a more fair estimate of the skill of the system.

Considering the RPS (black contour lines in Figure 2e–h), we can see an additional effect to that observed above. Note that for the definition of the tercile thresholds, we assumed stationary terciles in the left column (Figure 2e, g), but accounted for the changing climatology in the right column (Figure 2f, h). In the hindcast period, the changing thresholds have no effect on the RPS (same black contours in g and h). In the forecast period however, assuming stationary thresholds leads to the estimation of very low RPS values and a strong decrease of the score with increasing trends. Considering

changing terciles (Figure 2f) compensates for much of this effect. Although there is still a decrease of the RPS with the trend, this decrease is strongly reduced compared to the forecast RPS with stationary climatology but also compared to the hindcasts (steeper black lines in Figure 2e, g, h compared to f). When we consider the skill score (i.e., the RPSS, shading in Figure 2e–h), we can see that, as for the CRPSS, when we account for the changing climatology in the reference forecast the estimate of the skill comes to lie very close to the detrended skill of the system (compare shading in Figure 2e, g with f, h).

3.2.3 | Forecast skill inflation

We saw that, in the presence of a trend, the estimated skill of the synthetic prediction system depends on the definition of the climatology and can vary between hindcast and forecast period, despite the synthetic model having constant detrended skill. The skill increase that we described above can be entirely explained by the changing climatology. To measure the effect of assuming a stationary climatology, we define the inflation I as the difference between the skill estimated under the assumption of a stationary climatology $\mathrm{SKS}_{\mathrm{stat}}$ and the detrended skill of the system $\mathrm{SKS}_{\mathrm{detr}}$:

$$I = \mathrm{SKS}_{\mathrm{stat}} - \mathrm{SKS}_{\mathrm{detr}}. \tag{17}$$

The inflation is thus the difference between the shading in the left and the right columns of Figure 2 and is displayed in Figure 3. Here, it is readily evident that the changing climatology leads to an inflation of both the CRPSS and the RPSS with increasing trend. The inflation is much stronger in the forecast period (Figure 3a, c) than in the hindcast period (Figure 3b, d), which is mainly due to the fact that a stationary reference climatology is a poorer model for a period outside of the hindcast period when the climatology is in fact changing. The inflation is particularly severe when the model has little to no skill α while I increases much less with increasing trend when the detrended skill is high. For a forecast without any detrended skill, the forecast CRPSS increases by approximately 0.05 for a 5% increase in trend variance. The RPSS increases by 0.075 for a 5% increase in trend variance when the system has no detrended skill. However, it should be noted that the inflation as we define it in Equation (17) is an absolute measure. We can also consider the inflation relative to the actual system's skill by dividing I by SKS_{detr}. This relative inflation is mainly a function of the trend variance. We find that the forecast (hindcast) CRPSS is inflated by approximately 5% (3%) for a 5% increase in variance explained by the trend. The forecast (hindcast) RPSS inflates more strongly at approximately 8.5% (4%) for a 5% increase in trend variance. The increase of the inflation with trend is well approximated by a linear function (somewhat less so for the forecast CRPSS, not shown) for trends that explain less than 60% of the variance ($\sigma^2_\Delta < 0.6$).

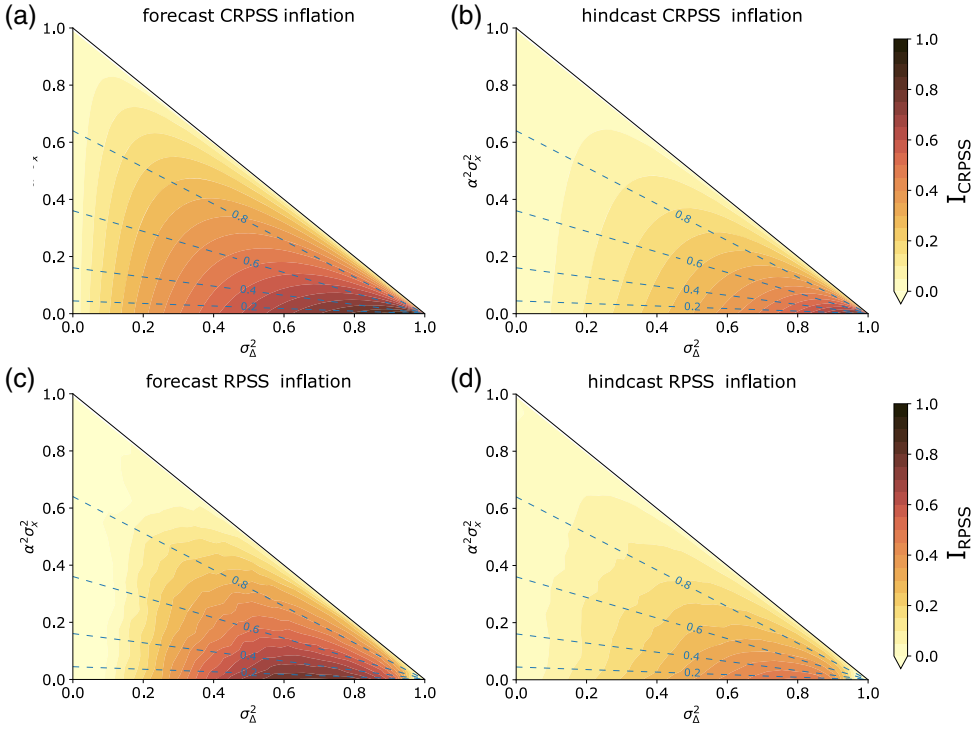

FIGURE 3 Skill inflation I as defined in Equation (17) in the synthetic model. Shading denotes the inflation I for the skill scores evaluated over the forecast and hindcast periods as indicated in the respective panel titles. Blue contours show the prescribed correlation α for reference [Colour figure can be viewed at wileyonlinelibrary.com]

When the prediction system reproduces the trend in the verification perfectly, the scores of both hindcasts and forecasts improve (corresponding to decreasing CRPS and RPS) the larger the trend is. This happens despite the fact that the model's skill at predicting the variability on the time-scale of interest (the detrended variability) does not change. The decisive factor for this inflation is the amount of variance explained by the trend σ^2_Δ during the hindcast period. An increase in σ^2_Δ leads to an increase in the sharpness of the prediction ensemble (i.e., a narrower ensemble distribution), while the reliability of the forecasts remains unaltered. For a categorical score like the RPS, this effect is further exacerbated when the categories' thresholds are considered constant over the hindcast and forecast periods. When computing the skill score for these forecasts with respect to a climatological reference forecast, the non-stationarity of the time series can be accounted for in the reference forecast to avoid inflation of the skill scores. Only if the reference captures the changing climatology does the skill score reflect the skill of the prediction system at forecasting variability on the time-scale of interest.

3.3 | Effects of mis-estimation of the trend in forecasts

Since we do not expect an actual prediction system to perfectly reproduce the observed trend, we test the sensitivity of the skill improvement to the error in the trend. Thus, we next allow the parameter p (Equation (15)) of our model to vary where $p < 1$ means an underestimation of the trend in the prediction system and $p > 1$ an overestimation. The inflation I as a function of the trend and the mis-estimation p are displayed in Figure 4. Note that the choice of p as a coordinate results in lower absolute errors $e_\Delta = \sigma^2_\tau - \sigma^2_\Delta$ occupying a larger area in the plot, which is evident from the grey contour lines that show e_Δ. For Figure 4, we use a fixed level of $\alpha = 0.4$, which is the approximate global average correlation skill for week 3 subseasonal 2 m temperature forecasts.

Along the horizontal grey line in Figure 4 ($p = 1$) we see the same increase of the skill score as in the respective panels of Figure 2. However, even when the trend is over- or underestimated (moving up or down, respectively, relative to the horizontal grey line) inflation is generally positive. In fact, the inflation is even stronger when strong trends are underestimated. The reason is that $\mathrm{SKS}_{\mathrm{detr}}$ decreases more rapidly than $\mathrm{SKS}_{\mathrm{stat}}$ with larger mis-estimation. This effect is present in both hindcasts (Figure 4b, d) and forecasts (a, c) but is much more pronounced in the forecasts. This contrast between hindcasts and forecasts can be explained by the fact that the forecasts exhibit a large unconditional bias when the trend is more strongly mis-estimated, while the hindcasts do not. In addition, as discussed above, a reference forecast with a stationary climatology has more skill for larger trends in the forecast period. The combination of the

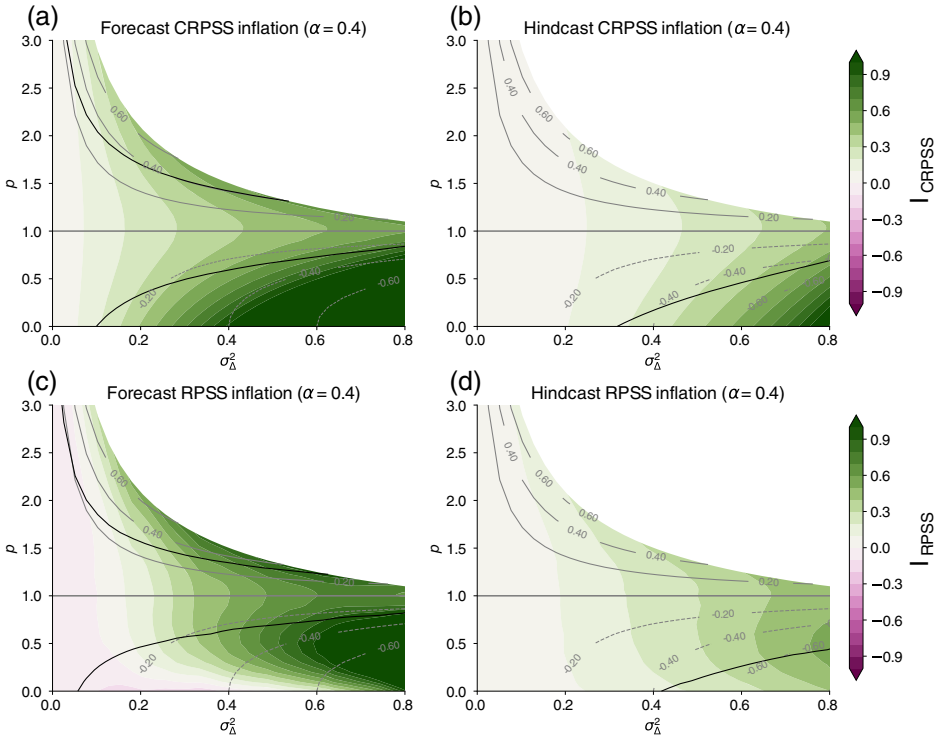

FIGURE 4 Dependence of the skill inflation I (Equation (17)) on the mis-estimation p of the trend by the forecasts for the synthetic model for a fixed level of detrended correlation skill $\alpha = 0.4$. I is shown for (a, b) the CRPSS and (c, d) the RPSS and for both (a, c) forecasts and (b, d) hindcasts. The shading shows I as a function of the fraction of variance of the verification contained in the trend (σ^2_Δ), and the relative error p of the forecasts at reproducing the trend. Grey contour lines indicate the absolute trend error e_Δ (see text). Black lines indicate where the non-inflated skill $\mathrm{SKS}_{\mathrm{detr}}$ is zero ($\mathrm{SKS}_{\mathrm{detr}}$ is positive between black contours) [Colour figure can be viewed at wileyonlinelibrary.com]

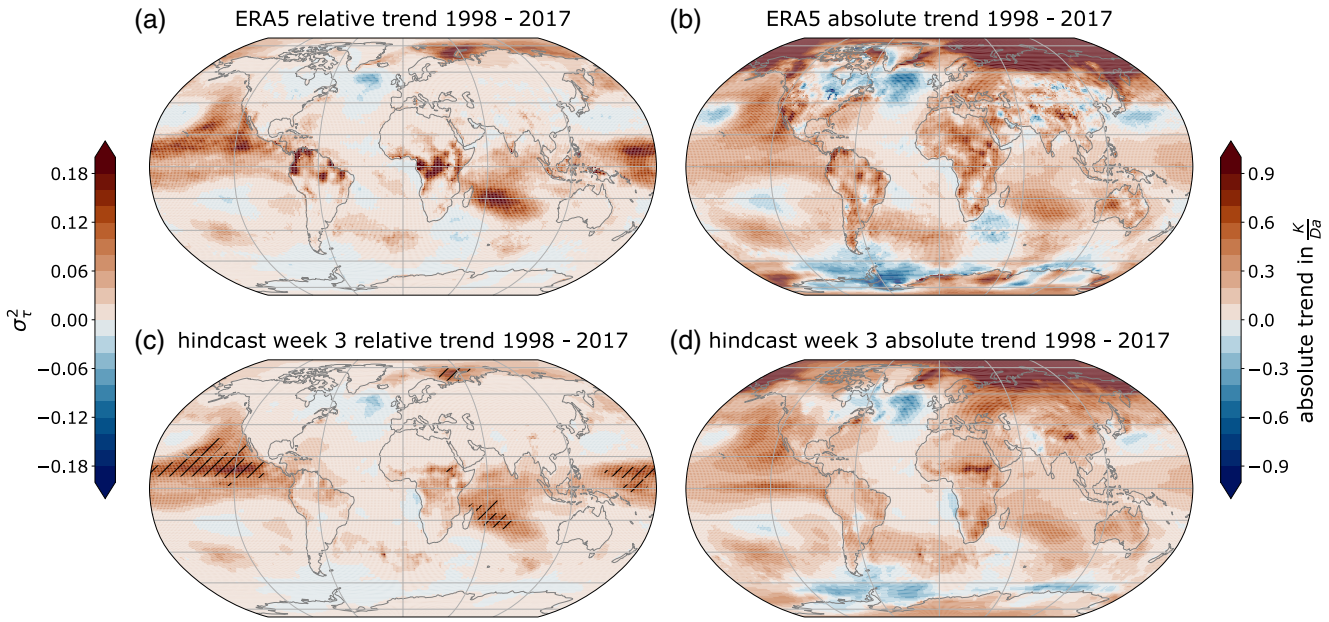

FIGURE 5 Trend (1998–2017) of 7-day mean T_{2m} in (a, b) ERA5 and (c, d) the week 3 hindcasts. In (a, c), the trend is shown as the fraction of variance it explains during the hindcast period (σ^2_Δ), multiplied by the sign of the trend. (b, d) show the absolute temperature trend (K per decade). Hatching in (a, c) shows where the trend explains more than 10% of the variance of the time series [Colour figure can be viewed at wileyonlinelibrary.com]

aforementioned factors results in the inflation being stronger in the forecast period and I having a tendency to become larger for deviations of p from 1 when the actual trend is strong.

In summary, our analyses show that the skill can be inflated in the presence of a trend even if it is not perfectly reproduced in the prediction system. The inflation is more severe in the forecast period and especially strong – in fact stronger than if the trend were perfectly reproduced – when large trends in the verification are underestimated by the prediction system. For weaker and perhaps more realistic trends, however, the inflation is fairly similar for all levels of mis-estimation.

4 | TREND EFFECT IN AN OPERATIONAL SUBSEASONAL PREDICTION SYSTEM

In order to compare to the synthetic forecasts from the example above (Section 3.3), we now analyse the behaviour of the ECMWF hindcasts. We first consider the trends of the verification (ERA5) over the hindcast period. The trends of the standardised and absolute anomalies are shown in Figure 5a, b, respectively. Over large parts of the globe the 20-year trends are generally in agreement with global warming signals computed over much longer periods (compare with IPCC, 2013, e.g., figure 2.21). For instance, strong absolute temperature trends (Figure 5b) are found in the Arctic regions, consistent with the well-known Arctic amplification signal (Screen and Simmonds, 2010). We can further see enhanced warming over Siberia, as well as generally stronger trends over land than over the ocean. There is also a pronounced lack of warming in the North Atlantic, consistent with the North Atlantic warming hole (Drijfhout *et al.*, 2012). Despite these consistencies there are also some differences to longer-term trends. Specifically, negative trends south of South America, the relatively strong warming over large parts of the tropical and eastern North Pacific Ocean and some smaller-scale trend patterns are not typically identified as long-term temperature change signals. The reason for this is that the available hindcast period (1998–2017) is not long enough to average out decadal to multi-decadal variability. The warming in the Pacific in Figure 5 is a manifestation of the Pacific Decadal Oscillation having mainly resided in its negative phase from the late 1990s until approximately 2015 and a subsequent switch to more positive conditions within the last couple of years (figure 1 in Newman *et al.*, 2016). Thus, although the linear trends computed from the hindcast period are not purely a manifestation of global warming, they do represent deviations of each grid point's temperature time series from stationarity during the hindcast period and can thus be treated approximately as the trends in our synthetic model.

In Section 3.2, we saw that the effect of the trend on the prediction skill is determined by the amount of a time series' variance that is explained by the trend. The absolute trends are thus not appropriate to identify regions where the forecast skill could be affected by inflation. Figure 5a shows the fraction of variance explained by the trends of 7-day mean T_{2m} from ERA5 in the hindcast period. Clearly, in many regions with strong absolute trends, the trends are actually small in comparison to the week-to-week variance and thus are not necessarily contributing strongly to the forecast skill. This is the case for large parts of the Arctic (except for the Barents Sea and parts of the Kara Sea) and many land areas. Instead, in terms of the variance explained by the trend, the oceans and tropical belt show rather stronger signals. The regions where the trend explains more than 10% of the variance are hatched in Figure 5a. Following the synthetic model from the previous section, the hindcast skill could be enhanced by up to 5% (not accounting for a mis-estimation of the trend) and the forecast skill could be enhanced even more strongly. We expect the skill to be most affected by the trend in these highlighted regions.

The temperature trends estimated from ERA5 are reasonably well reproduced by the forecast model (Figure 5c, d). On average, there is a weak tendency of the model to underestimate the trends, both in absolute terms and when considering the amount of variance they explain. However, there are substantial spatial variations in this mis-estimation, which is discussed further in Section 4.2. The fact that the mis-estimation is largely similar - independent of whether we consider the absolute trends or their contribution to the variance - indicates that the total variance in ERA5 is well reproduced in the forecast model (not shown). Although we only show the trend computed from the hindcast temperatures at week 3 to represent a subseasonal lead time, in any other forecast week the fraction of variance explained by the trend is almost the same as in week 3 (not shown).

4.1 | Comparison of the ECMWF system with the synthetic model

Knowing the trend patterns in the hindcast period and their differences in both the verification and the hindcasts, we now assess how the probabilistic skill of the operational prediction ensemble behaves in comparison to the synthetic ensemble from the previous section. For this, we sort the grid points of the model by the variance explained by the trend in the verification σ^2_Δ and the ratio p of the relative trend slopes in the hindcasts and the verification and bin the data into 100 bins in each of these dimensions (same dimensions as in Figure 4). Note that we exclude grid points where the sign of the trend disagrees between the model and reanalysis ($p < 0$), since we did not consider these cases in our synthetic model. We thus retain approximately 91% of all grid points. We then compute the bin-average of the inflation I of both hindcasts and forecasts as defined in Equation (17). These are shown for forecast week 3 in Figure 6. Focusing first on the inflation in the hindcast period (Figure 6b, d), we can see some agreement with the synthetic model (Figure 4b, d) for both the CRPSS and the RPSS. While the inflation is generally low for the hindcasts, some darker green bins are located where the trends are larger and $p < 1$, which indicates a similar tendency for I in the operational hindcasts as in the synthetic ones.

Inflation in the operational forecasts (Figure 6a, c) shows notably stronger deviations from the simple model (Figure 4a, c) than the hindcasts. For the forecast CRPSS (Figure 6a), inflation is mostly positive and shows a tendency to increase with stronger trends, much like in the synthetic model. However, there is also a notable area where I is negative. Negative values of I in the context of the synthetic model would imply a detrended skill that is higher than the skill estimated without accounting for the underlying trend and hence does not occur in the synthetic model for the forecast CRPSS. The reason for the occurrence of 'deflation' ($I < 0$) lies in the violation of the assumption that the trend estimated from the hindcast period is continued in the forecast period. In our synthetic model, this assumption is valid by design, which results in the climatological reference forecast that accounts for the trend being perfectly calibrated in both the hindcast and the forecast periods. Additionally, accounting for the trend results in a narrower spread of the climatology as compared to one estimated under the assumption of stationarity. Thus, the reference score in the synthetic model will always be better when accounting for the trend. In reality, the assumption does not hold, because a linear trend estimated from a period of 20 years is not a robust estimate of the actual long-term changes in the climate as they are affected by other low-frequency (e.g., multi-annual to multi-decadal) variability. As a result, the reference climatology including a trend can be poorly calibrated to the forecast period. Note, for instance, that a climatology forecast that includes the trend will produce a non-zero chance for the occurrence of previously unobserved temperatures. Especially when the long-term temperature tendency in the forecast period is actually weaker than estimated from the hindcast period or even reversed, values outside of the climatological range have a much reduced chance of occurrence. In those cases, a stationary climatology will be better calibrated during the forecast period than one assuming a trend, thus presenting a more competitive reference forecast, which results in $SKS_{stat} > SKS_{detr}$ and

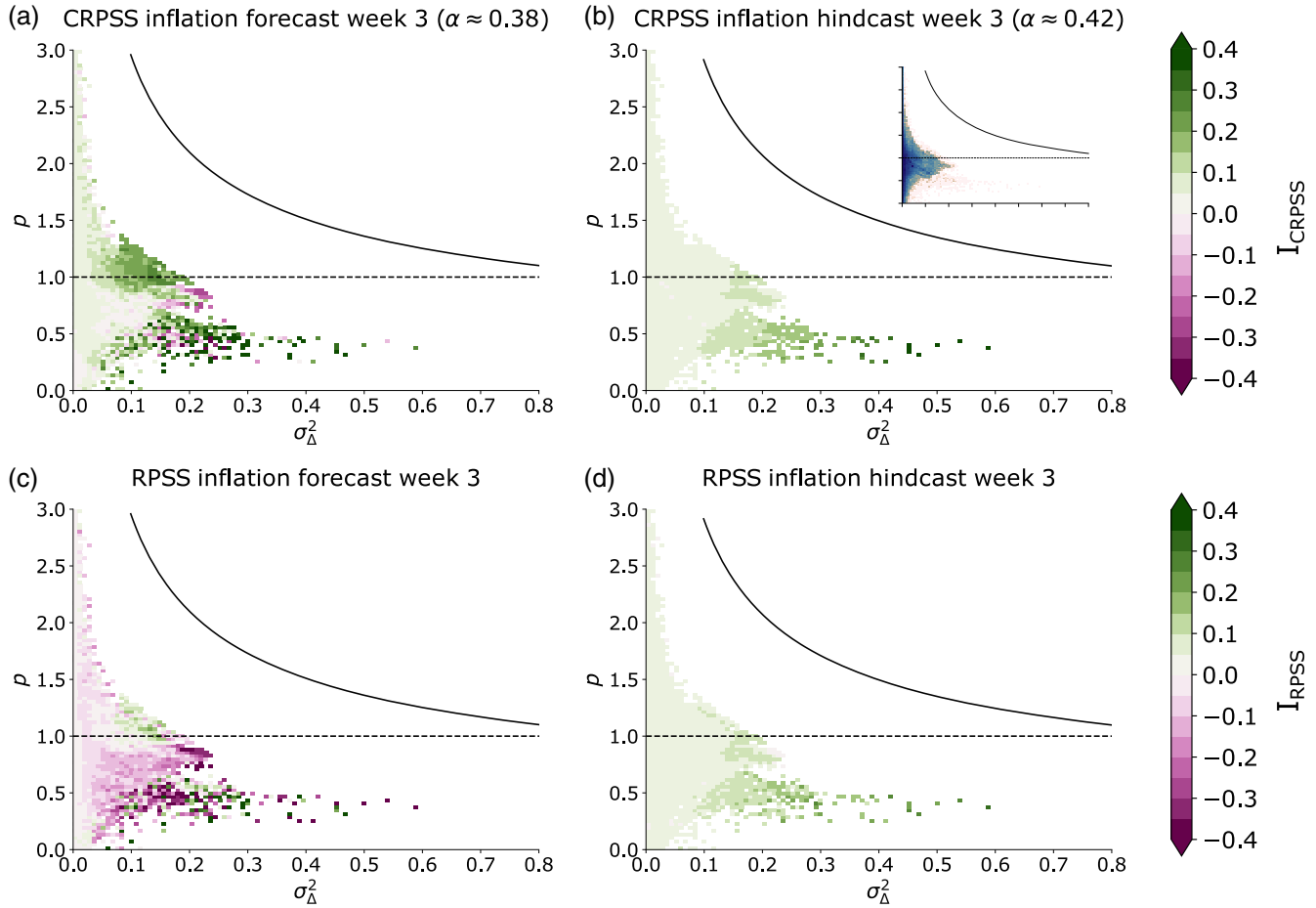

FIGURE 6 Inflation I as in Figure 4 but for week 3 forecasts and hindcasts of the subseasonal model. Note the different colour scale from Figure 4. Every grid point of the model was sorted into a bin according to the variance explained by the trend in the verification during the hindcast period (σ^2_Δ) and the factor p by which the trend is mis-estimated by the hindcasts. The grid point values of I are then averaged over 100×100 equally spaced bins between $0 \leq \sigma^2_\Delta < 1$ and $0 \leq p < 3$ (without weighting by grid-cell area). The inset in (b) shows the density of points in each bin on a logarithmic scale (valid for all panels). The black dashed line shows where the observed trend is perfectly reproduced by the forecasts. The black solid line shows the theoretical upper limit for p, which is $p_{mx} = \sigma^{-1}_\Delta \left(1 - \alpha^2 \sigma^2_x\right)$ where we use the average of the detrended correlations in the hindcasts (b, d) and forecasts (a, c) at all grid points as α [Colour figure can be viewed at wileyonlinelibrary.com]

consequently $I < 0$. We also used our synthetic model to test the effect of a trend that is different in the forecast period than in the hindcast period and found that in these cases it is possible to get negative inflation (not shown).

The inflation of the forecast RPSS (Figure 6c) is dominated by bins with $I < 0$ with only few bins indicating an actual inflation ($I > 0$). The reason for the occurrence of deflation is, as for the CRPSS, the poor representativeness of the estimated trend for future low-frequency changes in the forecast period. However, while the trend estimate enters only in the climatological reference forecast for the CRPSS, for the RPSS it also enters in the tercile definition, which makes the RPSS much more sensitive to inaccuracies in the climatology than the CRPSS, explaining the dominance of $I < 0$ for the forecast RPSS, while for the CRPSS I is still mostly positive. This stronger sensitivity of the RPSS can also be seen in the synthetic model when letting the trend in the forecast period be different from the trend in the hindcast period (not shown). As a result of this sensitivity, the estimated inflation of the RPSS in the operational prediction system during the forecast period is mainly negative.

In summary, the inflation in the hindcasts indicates that the synthetic model captures the effect of a trend over the hindcast and forecast periods fairly well. This confirms the general suitability of our synthetic model to measure the inflation due to a known long-term trend. However, since only a small part of the bins in the σ^2_Δ–p domain is populated (inset of Figure 6b) in the operational system, direct comparison with the synthetic model is rendered difficult. Additionally, while all hindcasts in Figure 4 have the same level of detrended skill ($\alpha = 0.4$) and the average skill over all grid points is close to 0.4, the forecasts at individual grid points naturally have different levels

of detrended skill. For the forecasts, there is an apparent inconsistency between the operational prediction system and the synthetic model, especially for the RPSS. This inconsistency is not necessarily due to the lack of realism of the synthetic model. It rather reflects the fact that the linear trends estimated from 20 years of data are poor representations of the actual long-term non-stationarity of the climate. The estimated trends are influenced by multi-annual to multi-decadal variability and thus do not realistically represent the linear change of the climatology in the forecast period. We thus conclude that our simple model is suitable to represent the effect of a long-term linear trend on the forecast skill, but the exact quantification of this inflation is hampered by the difficulty of robustly estimating the trend from a limited hindcast period.

4.2 | Geographical distribution of the trend effect in hindcasts

In addition to our above consideration of the inflation in the ECMWF system in coordinates that facilitate comparison with the synthetic model, we now consider the geographical distribution of the effect in the hindcasts. This serves the purpose of identifying regions where inflation could be a factor in the evaluation of subseasonal forecast skill. In the previous section, we saw that there is negative inflation in the forecast period, mainly as a result of mismatches between the estimated trends and the long-term temperature tendency during the forecast period. To analyze the potential inflation due to trends in the data, the hindcast period is thus more suitable, since the trend represents the optimal linear fit to the temperatures in this period.

Figure 7a, b show the inflation of the week 3 hindcast CRPSS and RPSS, respectively. The skill inflation is positive throughout most of the globe and consistent between CRPSS and RPSS with the latter appearing slightly more noisy. Generally, the magnitude of I is small ($I < 0.1$), but given the low skill of the forecasts for week 3 or longer lead times, inflation can regionally be a decisive factor in whether a forecast has skill or not. Figure 7a, b indicate strongest inflation in the eastern North Pacific, the southern Indian Ocean, western and central Africa, and equatorial South America. Substantial inflation also occurs in the central to eastern tropical Pacific, which is particularly interesting in the light of the unresolved future changes in ENSO under climate change, e.g., Heede *et al.*, (2020). Although slightly smaller in amplitude compared to the aforementioned regions, part of the Barents and Kara Seas appears as another regional maximum of inflation. All these regions are highly consistent with the strongest positive relative temperature trends over the hindcast period (Figure 5a). As was already indicated in Section 4.1, the mis-estimation p (Figure 7c) of the trend plays only a secondary role. Comparing Figure 7c with Figure 7a, b, it becomes clear that inflation occurs predominantly where the ERA5 trend is well reproduced ($p \approx 1$, no shading) or underestimated ($p < 1$, blue shading), although it should also be noted that areas of $p > 1$ are generally less abundant, and even entirely absent for strong trends (also Figure 6). In summary, skill inflation due to the presence of a trend in the hindcast period occurs to varying degrees globally with the largest magnitudes where relative trends are strongest during the hindcast period, which is predominantly the case in the Tropics.

5 | DISCUSSION

We have used a simple synthetic forecast model to show that probabilistic forecast skill can be enhanced in the presence of a trend. Especially in the case of categorical forecasts, it is already visible from a simple illustration like Figure 1 why a trend in the verification period can have an effect on the predicted category and thus on the skill score: if the categories are defined as percentiles of the climatological distribution and estimated under the assumption of a stationary climate in the hindcast period, for a positive trend the upper category is more likely to occur in the forecast period. A prediction system that knows only this trend but has no skill at predicting variability on the time-scale of interest will appear to have skill since its (random) forecasts also lie in the upper category more often. This concept was illustrated and discussed before by Livezey, (1999, specifically their figure 10.2). Our results confirm Livezey's arguments quantitatively for a categorical skill score, namely the RPSS. We further show that there is an enhancement of the probabilistic skill for continuous forecasts (namely, the CRPSS) which do not rely on the definition of categories that are estimated from the hindcast period. We refer to the aforementioned enhancement of the skill scores in the presence of a long-term trend as inflation. While a prediction system is correct in forecasting higher temperatures more frequently when there is a positive temperature trend, it is eventually up to the forecast user to decide whether the skill that arises from the trend justifies the use of a computationally expensive dynamical model any more than the skill arising from reproducing the seasonal cycle does. After all, a trend can be understood as a shift or non-stationarity in the mean of the climatological distribution and can be estimated reasonably well from a sufficiently long past period.

The issue of skill inflation by not accounting for varying climatologies has also been addressed by Hamill and

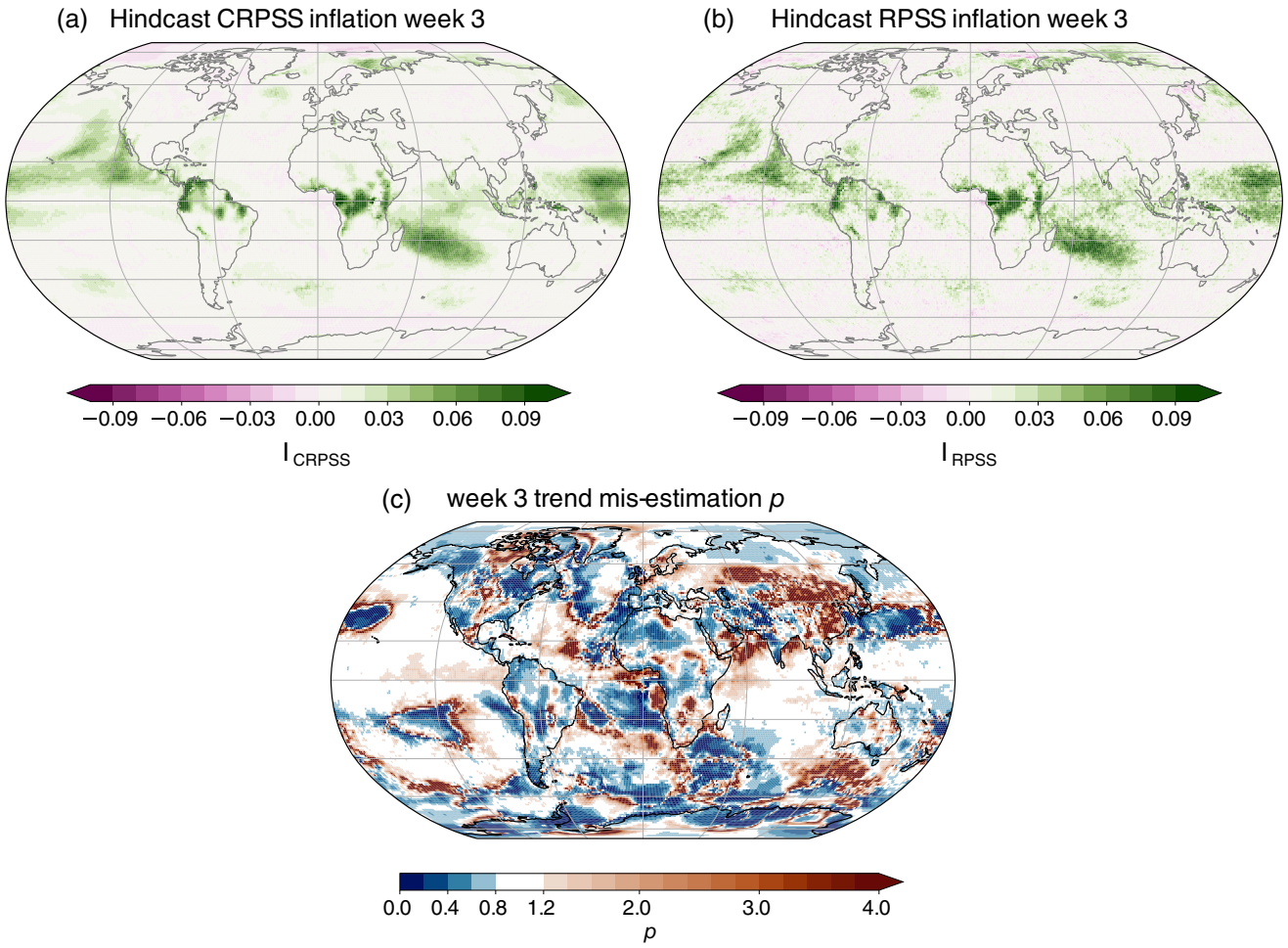

FIGURE 7 Inflation I of the (a) CRPSS and (b) RPSS of the ECMWF week 3 hindcasts of 7-day mean, standardised 2 m temperature anomalies at every grid-point. (c) shows the mis-estimation factor p between the temperature trend during the hindcast period in ERA5 and in the model [Colour figure can be viewed at wileyonlinelibrary.com]

Juras (2006). They showed that categorical skill scores (both deterministic and probabilistic) can appear higher when locations with different climatologies are pooled to estimate the climatological reference as opposed to accounting for the different event frequencies when scoring the forecasts. Our results are in line with this finding, the difference being mainly that we do not consider spatially varying climatologies but a monotonously changing (and thus non-stationary) climatology over the prediction period. The effect of this monotonous change is especially obvious for the RPSS; in case the trend is not accounted for, the average RPSS increases when evaluating a longer forecast period, which is at odds with our understanding of skill. This problem is in fact one of the reasons why the subseasonal hindcasts at the ECMWF are computed for only 20 years instead of 30–40 years, which is a common hindcast period for seasonal forecasts.

The hindcasts of the operational prediction system, on average, broadly follow the behaviour of the synthetic prediction system that we designed in terms of the effect of a trend on the probabilistic skill. This is despite the fact that the toy model represents a strong simplification of the statistical properties of the operational system. Here, we would like to point out what we consider the four main limitations on the realism of the synthetic model that likely play a role in causing some of the differences with respect to the operational prediction system. The limitations are largely coincident with those discussed by Weigel *et al.*, (2008) who used a similar set-up of synthetic forecast–verification pairs.

1. The verification is drawn from a normal distribution at every time step. Although this assumption might hold in reality for weekly means of standardised 2 m temperature in many regions of the globe, it is certainly violated in places where temperatures are subject to feedbacks at certain times of the year. For instance, this could be the case for ocean regions with sea ice during

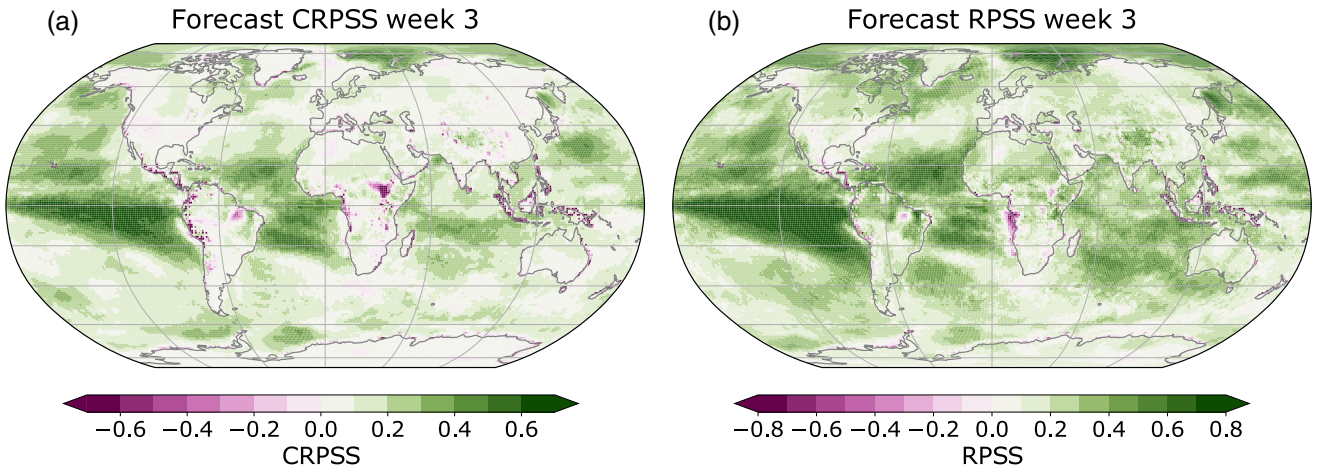

FIGURE 8 Corrected skill of the ECMWF system over a climatological reference with linearly changing mean. (a) shows the CRPSS for week 3 forecasts initialised between 1 January 2018 and 31 December 2020 for 7-day mean, standardised temperatures with ERA5 as verification. (b) shows the respective RPSS for tercile forecasts. Both the CRPSS and RPSS were computed with respect to a reference that accounts for the trend in ERA5 [Colour figure can be viewed at wileyonlinelibrary.com]

parts of the year, land regions where snowmelt occurs or places in which land–atmosphere feedbacks tend to be relevant.

2. Just as the synthetic verification follows a normal distribution, so do the synthetic predictions. This leads to almost perfect calibration, that is, an optimal reliability component of the CRPS and RPS. Since we apply a rather crude calibration to the operational predictions, the reliability is certainly worse than optimal. This affects the considered scores (Ferro *et al.*, 2008) and we hypothesise that the less-than-perfect calibration is responsible for parts of the deviations between Figures 6 and 4.
3. The synthetic model has stationary skill. While we do prescribe a change in the climatology, the skill of the synthetic model does not change over the simulated period. This assumption is likely violated in the operational forecasts, as has been indicated for instance for seasonal forecast skill for the North Atlantic Oscillation (Weisheimer *et al.*, 2017). To what degree this affects the 23-year period that we consider is difficult to quantify and could vary considerably between regions.
4. The synthetic model only considers a linear change in the mean of the climatology. Non-stationarity can be assumed to be more complex and affect other moments of the climatological distribution (e.g., Schär *et al.*, 2004). Related to this, the real trend itself likely exhibits seasonal variations.

Regardless of these limitations, the synthetic model describes the average behaviour of the hindcasts of the ECMWF prediction system quite well.

Despite the simplicity of the synthetic model, its consistency with the operational hindcasts indicates that it is suitable to represent the potential inflation of the skill that can be introduced by a long-term trend over the joint hindcast and forecast periods. We also showed that this consistency breaks down in the forecast period. Part of this mismatch could be due the aforementioned lack of realism of the synthetic model. However, we also discussed in Section 4.1 that the trend estimated from the hindcast period can be a poor representation of actual long-term climatological changes over the forecast period (also Livezey *et al.*, 2007; Wilks, 2013; Wilks and Livezey, 2013). Further analysis confirms that – rather than a lack of realism of the simple model – this poor suitability of the trend to represent temperature changes during the forecast periods is in fact the major reason for the inconsistency between the synthetic model and the operational forecasts (Appendix S1).

Furthermore, in the synthetic model, we can separate the trend clearly from all other variability. In reality, even if we knew there were in fact some underlying, perfectly linear long-term change in the temperature climatology, correctly estimating this change from a time series of only 20 years would be strongly hampered by the presence of other low-frequency variations. A longer observational record would provide a more robust estimate of the trend in the verification, which could be used to get an estimate of the potential skill inflation using Figure 3. Whether the models reproduce this trend, however, cannot be clearly determined from the more limited hindcast period but matters for the strength of the inflation. This makes a realistic estimation of the inflation in any prediction system difficult.

Finally, keeping in mind that the estimation of the inflation effect is made difficult by the fact that the trend cannot be robustly estimated from the 20-year hindcast period, we compute the corrected global skill estimates for ECMWF week 3 forecasts initialised between 2018 and 2020. Figure 8 shows the CRPSS and RPSS for the 7-day mean, standardised temperatures with ERA5 as verification, indicating where the forecasts have skill over a climatological forecast that accounts for linear changes in the mean. The overall pattern is very similar between the CRPSS and the RPSS, with generally higher skill over the ocean and in the Tropics. The reason for the enhanced skill over the ocean is the longer persistence of ocean anomalies. In the Tropics, the atmosphere additionally exhibits strong coupling with the ocean, which is the major cause for the extended predictability of phenomena like the ENSO and the MJO. The remote influence of these highly predictable phenomena is likely responsible for parts of the enhanced week 3 skill in extratropical regions like the North Pacific and west of Australia. Week 3 forecast skill is also high in the Barents and Kara Seas but note that the observed large and potentially nonlinear long-term temperature changes in this region likely have an effect on the forecast skill that goes beyond the inflation we accounted for (Section 4.2).

6 | CONCLUSION

A trend in a time series represents a non-stationary component in the climatology. We have shown that a simple linear trend can improve the forecast skill of a toy forecast even in the absence of any actual predictive skill on the time-scale of interest. This skill 'inflation' is a function of variance explained by the trend but also of the mis-estimation of the trend in the prediction system. The effect is present both in skill estimates based on categorical scores (here, the RPSS) and continuous scores (CRPSS). The effect is stronger for the RPSS and is further enhanced when averaging the RPSS over a longer forecast period. The reason for the strong sensitivity of the categorical skill is that the estimate of the climatology enters both in the reference forecast and in the determination of the category thresholds. In a simple synthetic model that simulates the overall statistical properties of an ensemble forecasting system, the forecast (hindcast) CRPSS is enhanced by 5% (3%) for a 5% increase in trend variance while the forecast (hindcast) RPSS is enhanced by 8.5% (4%) per 5% in the chosen set-up. The inflation can become even stronger if the trend is mis-estimated in the synthetic forecasts relative to the trend in the verification. The inflation simulated by the synthetic model is in good agreement with the average effect in the hindcasts of an operational prediction system (the ECMWF extended-range forecasting system). In the forecast period, the estimation of the inflation is strongly hampered by the difficulty to robustly estimate the true trend from a limited hindcast period of 20 years and the influence of other low-frequency variations on the estimated trend. The inflation is strongest in the Tropics where the trend accounts for a larger part of the variability than in other regions, which is consistent with the simple model. Even though we here focus on subseasonal forecasts, our results can be generalised to some degree to forecasts at any time-scale and lead time. The trend effect is a function of the signal-to-noise ratio, that is, the trend relative to the internal variability on the considered time-scale. Thus, considering for instance decadal forecasts where annual means are predicted, we would expect to see a stronger effect of the trend than for the prediction of weekly means in subseasonal forecasts, mostly because interannual temperature variance is expected to be lower than weekly variance, which results in a larger fraction of total variance being explained by the long-term trend. Similarly, spatial aggregation has an effect on the signal-to-noise ratio and thus the effect on skill scores of regional averages will be different than shown here, even when the same time-scale is considered, since the trend could make up for larger parts of the variance of spatially aggregated data. The results of our synthetic model allow for a simple benchmark estimate of the contribution of a (known) long-term trend on the skill of a forecast. It can be argued that this potential increase needs to be accounted for when reporting the skill of a forecast. We here call this skill increase inflation since – similar to the seasonal cycle, which is usually subtracted before computing the skill – the trend can be estimated from a sufficiently long observational record without the need for a dynamical prediction system. In reality, however, it is not straightforward to accurately quantify the inflation of the forecast skill because the trend can often not be estimated robustly from the available hindcast periods. We nevertheless think that forecasters should be aware of the potential effect that a changing climatology can have on the forecast skill because clearly communicating where the forecast skill stems from will enhance the users' confidence in forecast products.

ACKNOWLEDGEMENTS

We thank the ECMWF for providing extended-range forecasts from the S2S prediction project through the MARS archive (https://apps.ecmwf.int/datasets/data/s2s/) and Copernicus for providing the ERA5 data through the Climate Data Store (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview; both accessed 4 March 2022). All analyses and visualisations were carried out using the Python packages

numpy, matplotlib, cmcrameri, scipy and properscoring. Funding from the Swiss National Science Foundation to O.W. and D.D. through project PP00P2_170523 is gratefully acknowledged. The authors thank two anonymous reviewers for their insightful comments during the review process. We would also like to thank Antje Weisheimer and Christof Appenzeller for an interesting discussion on the results of this study. Open Access Funding provided by Eidgenossische Technische Hochschule Zurich. [Correction added on 23 May 2022, after first online publication: CSAL funding statement has been added.]

AUTHOR CONTRIBUTIONS

C. Ole Wulff: conceptualization; data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft; writing – review and editing. **Frédéric Vitart:** writing – review and editing. **Daniela I. V. Domeisen:** conceptualization; funding acquisition; project administration; resources; supervision; writing – review and editing.

ORCID

C. Ole Wulff https://orcid.org/0000-0001-7154-4812
Frédéric Vitart https://orcid.org/0000-0001-8485-7981
Daniela I. V. Domeisen https://orcid.org/0000-0002-1463-929X

REFERENCES

Alvarez, M.S., Coelho, C.A.S., Osman, M., Firpo, M.Â.F. and Vera, C.S. (2020) Assessment of ECMWF subseasonal temperature predictions for an anomalously cold week followed by an anomalously warm week in central and southeastern south America during July 2017. *Weather and Forecasting*, 35, 1871–1889.

Boer, G.J. (2009) Climate trends in a seasonal forecasting system. *Atmosphere–Ocean*, 47, 123–138.

DelSole, T. and Tippett, M.K. (2018) Predictability in a changing climate. *Climate Dynamics*, 51, 531–545.

Doblas-Reyes, F.J., Hagedorn, R., Palmer, T.N. and Morcrette, J.-J. (2006) Impact of increasing greenhouse gas concentrations in seasonal ensemble forecasts. *Geophysical Research Letters*, 33, 1–5.

Domeisen, D.I.V., Butler, A.H., Charlton-Perez, A.J., Ayarzagüena, B., Baldwin, M.P., Dunn-Sigouin, E., Furtado, J.C., Garfinkel, C.I., Hitchcock, P., Karpechko, A.Y., Kim, H., Knight, J., Lang, A.L., Lim, E., Marshall, A., Roff, G., Schwartz, C., Simpson, I.R., Son, S. and Taguchi, M. (2020) The role of the stratosphere in subseasonal to seasonal prediction: 2. Predictability arising from stratosphere–troposphere coupling. *Journal of Geophysical Research; Atmospheres*, 125(2). https://doi.org/10.1029/2019JD030923.

Drijfhout, S., van Oldenborgh, G.J. and Cimatoribus, A. (2012) Is a decline of AMOC causing the warming hole above the North Atlantic in observed and modeled warming patterns?. *Journal of Climate*, 25, 8373–8379.

Ferro, C.A., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Quarterly Journal of the Royal Meteorological Society*, 15, 19–24.

Hamill, T.M. and Juras, J. (2006) Measuring forecast skill: is it real skill or is it the varying climatology?. *Quarterly Journal of the Royal Meteorological Society*, 132, 2905–2923.

Heede, U.K., Fedorov, A.V. and Burls, N.J. (2020) Time scales and mechanisms for the tropical Pacific response to global warming: a tug of war between the ocean thermostat and weaker Walker. *Journal of Climate*, 33, 6101–6118.

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

Hoskins, B.J. (2013) The potential for skill across the range of the seamless weather–climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139, 573–584.

IPCC (2013). 2. Observations: Atmosphere and Surface, pp. 159–254 in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P. (eds). Cambridge, UK and New York, NY: Cambridge University Press.

Johnson, N.C., Collins, D.C., Feldstein, S.B., L'Heureux, M.L. and Riddle, E.E. (2014) Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Weather and Forecasting*, 29, 23–23.

Jolliffe, I.T. and Stephenson, D.B. (2012) *Forecast Verification: a Practitioner's Guide in Atmospheric Science* (2nd ed.). Chichester, UK: Wiley.

Jung, T., Kasper, M.A., Semmler, T. and Serrar, S. (2014) Arctic influence on subseasonal midlatitude prediction. *Geophysical Research Letters*, 41, 3676–3680.

Koster, R.D., Mahanama, S.P.P., Yamada, T.J., Balsamo, G., Berg, A.A., Boisserie, M., Dirmeyer, P.A., Doblas-Reyes, F.J., Drewitt, G., Gordon, C.T., Guo, Z., Jeong, J.-H., Lee, W.-S., Li, Z., Luo, L., Malyshev, S., Merryfield, W.J., Seneviratne, S.I., Stanelle, T., van den Hurk, B.J.J.M., Vitart, F. and Wood, E.F. (2011) The second phase of the global land–atmosphere coupling experiment: soil moisture contributions to subseasonal forecast skill. *Journal of Hydrometeorology*, 12, 805–822.

Lee, R.W., Woolnough, S.J., Charlton-Perez, A.J. and Vitart, F. (2019) ENSO modulation of MJO teleconnections to the North Atlantic and Europe. *Geophysical Research Letters*, 46, 13535–13545.

Leutbecher, M. (2019) Ensemble size: how suboptimal is less than infinity?. *Quarterly Journal of the Royal Meteorological Society*, 145, 107–128.

Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539.

Liniger, M.A., Mathis, H., Appenzeller, C. and Doblas-Reyes, F.J. (2007) Realistic greenhouse gas forcing and seasonal forecasts. *Geophysical Research Letters*, 34, 1–5.

Livezey, R.E. (1999). The evaluation of forecasts, pp. 179–198 in Analysis of Climate Variability, von Storch, H., Navarra, A. (eds). Berlin: Springer-Verlag.

Livezey, R.E., Vinnikov, K.Y., Timofeyeva, M.M., Tinker, R. and van den Dool, H.M. (2007) Estimation and extrapolation of climate normals and climatic trends. *Journal of Applied Meteorology and Climatology*, 46, 1759–1776.

Manrique-Suñén, A., Gonzalez-Reviriego, N., Torralba, V., Cortesi, N. and Doblas-Reyes, F.J. (2020) Choices in the verification of S2S forecasts and their implications for climate services. *Monthly Weather Review*, 148, 3995–4008.

Materia, S., Muñoz, Á.G., Álvarez-Castro, M.C., Mason, S.J., Vitart, F. and Gualdi, S. (2020) Multimodel subseasonal forecasts of spring cold spells: potential value for the hazelnut agribusiness. *Weather and Forecasting*, 35, 237–254.

Meehl, G.A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J.F., Stouffer, R.J. and Taylor, K.E. (2007) The WCRP CMIP3 multimodel dataset: a new era in climatic change research. *Bulletin of the American Meteorological Society*, 88, 1383–1394.

Müller, W.A., Appenzeller, C., Doblas-Reyes, F.J. and Liniger, M.A. (2005) A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, 18, 1513–1523.

Newman, M., Alexander, M.A., Ault, T.R., Cobb, K.M., Deser, C., Lorenzo, E.D., Mantua, N.J., Miller, A.J., Minobe, S., Nakamura, H., Schneider, N., Vimont, D.J., Phillips, A.S., Scott, J.D. and Smith, C.A. (2016) The Pacific decadal oscillation, revisited. *Journal of Climate*, 29, 4399–4427.

Peng, P., Kumar, A., Halpert, M.S. and Barnston, A.G. (2012) An analysis of CPC's operational 0.5-month lead seasonal outlooks. *Weather and Forecasting*, 27, 898–917.

Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127, 2473–2489.

Robertson, A.W., Vitart, F. and Camargo, S.J. (2020) Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *Journal of Geophysical Research: Atmospheres*, 125(6). https://doi.org/10.1029/2018JD029375.

Schär, C., Vidale, P.L., Lüthi, D., Frei, C., Häberli, C., Liniger, M.A. and Appenzeller, C. (2004) The role of increasing temperature variability in European summer heatwaves. *Nature*, 427, 332–336.

Screen, J.A. and Simmonds, I. (2010) The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature*, 464, 1334–1337.

Shukla, J., Anderson, J., Baumhefner, D., Brankovic, C., Chang, Y., Kalnay, E., Marx, L., Palmer, T.N., Paolino, D., Ploshay, J., Schubert, S., Straus, D., Suarez, M. and Tribbia, J. (2000) Dynamical seasonal prediction. *Bulletin of the American Meteorological Society*, 81, 2593–2606.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H., Hodgson, J., Kang, H.S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A.W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.J., Xiao, H., Zaripov, R. and Zhang, L. (2017) The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, 98, 163–173.

Vitart, F., Balsamo, G., Bidlot, J.-R., Lang, S., Tsonevsky, I., Richardson, D.S. and Balmaseda, M. (2019) ERA5 reanalysis used to initialise re-forecasts. *ECMWF Newsletter*, 161, 26–31.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2007) The discrete Brier and ranked probability skill scores. *Monthly Weather Review*, 135, 118–124.

Weigel, A.P., Liniger, M.A. and Appenzeller, C. (2008) Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?. *Quarterly Journal of the Royal Meteorological Society*, 134, 241–260.

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D.A. and Palmer, T. (2017) Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic Oscillation (NAO) and their potential value for extreme event attribution. *Quarterly Journal of the Royal Meteorological Society*, 143, 917–926.

White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A.P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K.V., Holbrook, N.J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T.J., Street, R., Jones, L., Remenyi, T.A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B. and Zebiak, S.E. (2017) Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, 24, 315–325.

Wilks, D.S. (2013) Projecting "normals" in a nonstationary climate. *Journal of Applied Meteorology and Climatology*, 52, 289–302.

Wilks, D.S. (2019). Forecast verification, pp. 369–483 in Statistical Methods in the Atmospheric Sciences. Amsterdam, Netherlands: Elsevier.

Wilks, D.S. and Livezey, R.E. (2013) Performance of alternative "normals" for tracking climate changes, using homogenized and nonhomogenized seasonal U.S. surface temperatures. *Journal of Applied Meteorology and Climatology*, 52, 1677–1687.

SUPPORTING INFORMATION

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

How to cite this article: Wulff, C.O., Vitart, F. & Domeisen, D.I.V. (2022) Influence of trends on subseasonal temperature prediction skill. *Quarterly Journal of the Royal Meteorological Society*, 148(744), 1280–1299. Available from: https://doi.org/10.1002/qj.4259