

A Technique for Seamless Forecast Construction and Validation from Weather to Monthly Time Scales

PAUL A. DIRMEYER

Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia

TRENT W. FORD

Illinois State Water Survey, Prairie Research Institute, University of Illinois at Urbana–Champaign, Urbana, Illinois

(Manuscript received 21 March 2019, in final form 21 April 2020)

ABSTRACT


Seamless prediction means bridging discrete short-term weather forecasts valid at a specific time and time-averaged forecasts at longer periods. Subseasonal predictions span this time range and must contend with this transition. Seamless forecasts and seamless validation methods go hand-in-hand. Time-averaged forecasts often feature a verification window that widens in time with growing forecast leads. Ideally, a smooth transition across daily to monthly time scales would provide true seamlessness—a generalized approach is presented here to accomplish this. We discuss prior attempts to achieve this transition with individual weighting functions before presenting the two-parameter Hill equation as a general weighting function to blend discrete and time-averaged forecasts, achieving seamlessness. The Hill equation can be tuned to specify the lead time at which the discrete forecast loses dominance to time-averaged forecasts, as well as the swiftness of the transition with lead time. For this application, discrete forecasts are defined at any lead time using a Kronecker delta weighting, and any time-averaged weighting approach can be used at longer leads. Time-averaged weighting functions whose averaging window widens with lead time are used. Example applications are shown for deterministic and ensemble forecasts and validation and a variety of validation metrics, along with sensitivities to parameter choices and a discussion of caveats. This technique aims to counterbalance the natural increase in uncertainty with forecast lead. It is not meant to construct forecasts with the highest skill, but to construct forecasts with the highest utility across time scales from weather to subseasonal in a single seamless product.

1. Introduction

The notion of seamless weather-to-climate prediction as a unified numerical modeling problem has been with us for at least a decade (WCRP 2005; Shukla et al. 2009; Shapiro et al. 2010; Brunet et al. 2010). However, there is an unescapable transition between weather forecasting as a deterministic initial-value problem to the probabilistic nature of longer-term time-averaged forecasts necessitated by the nonlinear character of the atmosphere–land–ocean system. Weather predictions validate on the shortest time interval of reported forecasts, which we define as “discrete” forecasts. These are typically hourly for nowcasting

and short-term weather forecasting, or as daily accumulations and averages out to as much as 2 weeks. These give way to longer-term means issued as “outlooks” by operational forecast centers on monthly or seasonal intervals. In fact, the coverage of the spectrum of time scales is not uniform, and prediction has focused on very specific ranges frequently tied to the calendar (Hoskins 2013).

In the middle of this dichotomy between discrete forecasts (and their validation) and longer-lead probabilistic time averages (and their validation) lies the so-called subseasonal-to-seasonal or S2S range of time scales. Recent efforts have been focused on understanding predictability and improving prediction skill in the S2S time range while endeavoring to use the same model configuration from weather to seasonal time scales (Vitart et al. 2017; Mariotti et al. 2018; Pegion et al. 2019). There has been much hope to extend weather forecast skill to weeks 3–4, but promising

 Denotes content that is immediately available upon publication as open access.

Corresponding author: Paul A. Dirmeyer, pdirmeye@gmu.edu

DOI: 10.1175/MWR-D-19-0076.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

prospects have been coming along slowly (Newman et al. 2003; Vitart 2014). Meanwhile, seasonal climate forecasts such as the outlooks issued by the Climate Prediction Center of NOAA, for example, begin with a 0.5–1.5-month lead-time average and include monthly and 3-month means that step forward in monthly intervals (Saha et al. 2014). This results in quite a prominent “seam” in the range where seamless forecasts have been envisioned.

Forecast uncertainty necessarily grows with lead time, so the characteristics of a forecast change markedly in the transition from weather to S2S time scales (Ebert et al. 2013). A key characteristic of the subseasonal transition is that uncertainty, and thus skill, is a function of the averaging period, both its width in days and the lead time of the forecast. One way to define a “seamless forecast” is one where the uncertainty is relatively constant with lead time, or at least ameliorated relative to the rapid decay of discrete deterministic forecasts verifying at a single point in time. One way to achieve that is to define the forecast, and its validation, over an averaging window that changes with lead time. This has many repercussions for forecast characteristics like error growth and saturation (Buizza et al. 2015), but follows our expectations of forecast predictability.

Imagine a situation where a brief but severe cold snap is predicted to occur 7 days after the initial time of a forecast, with unremarkable temperatures preceding and following the event. The event transpires much as it was predicted, except that the cold day arrived on the eighth day instead of the seventh day. A classical deterministic weather forecast validation, which is discrete in time (at least to the resolution of individual days), would severely punish the forecast for its 1-day timing error. Nevertheless, the forecast’s usefulness to interests such as farmers, electric utilities, first responders, and others who can use the intervening period to prepare is hardly impacted by the relatively small timing error. On the other hand, a time average forecast such as weekly mean temperatures would likely obscure the brief but extreme event hampering preparations. At that time scale, a 3-day mean could capture the severity and timing of the event with useful accuracy. It is easy to imagine similar situations for heat or precipitation, and at a variety of lead times.

Zhu et al. (2014) were among the first to concoct a sliding window, in which the validation of precipitation forecasts was performed over a window equal in width to the forecast lead time. More recently, a weighting scheme in time based on a Poisson function has been used to transition from a sharply peaked short-term average of forecast and validation data to a widening Gaussian distribution at time scales beyond a week or

two (Ford et al. 2018; Dirmeyer et al. 2018). This approach, applied to heat wave forecasts in the former case and general meteorological variables like temperature, humidity, and precipitation in the latter case, have suggested a way forward to a smooth and seamless transition between weather and climate time scales.

However, there is utility for discrete weather forecasts in the classical sense from 0 out to some number of days N , whose value depends on many factors but ultimately boils down to the duration of useful skill in the forecast. Ideal seamlessness would involve a way to maintain such classical weather forecasts for some period beyond the forecast initial date, and then seamlessly transition them to time averages so as to maintain useful skill and slow the growth of uncertainty with lead time. Within such a framework, the probabilistic attributes afforded by ensemble forecasts should also be fully accommodated.

In this study, we have refined the technique described in (Ford et al. 2018) to retain a discrete forecast approach over an arbitrary period of short-term forecasts, and allow for a controlled transition to time-mean forecasts. Section 2 examines the conceptualization with several approaches to the problem. Specific applications are demonstrated in section 3, including ensemble forecasts and validation. A discussion including caveats and considerations regarding parameter selection for the scheme are presented in section 4.

2. Weighting schemes

There are two aspects to seamless S2S forecasting. One is the retention of different treatments at short and long leads (i.e., discrete forecasts at short lead times), and some form of time average at longer lead times. The second, the key to seamlessness, is determination of a way to transition smoothly from one to the other as forecast lead increases. An ideal approach would have all of the characteristics in these two aspects.

For our examples, we define the shortest time scale, that for the discrete forecasts, as daily, thus avoiding fluctuations within the diurnal cycle. Daily means or totals are the basic increment of data considered, except for temperature extrema (maxima and minima) which can be traditional instantaneous values that are discretized at a daily interval. This corresponds well with much model forecast and hindcast data such as that from the international S2S prediction project database (Vitart et al. 2017) or the Subseasonal Experiment (SubX; Pegion et al. 2019). The validation day for any forecast is defined as a lead time in days from the initial state (τ). The forecast may represent the state only at day τ for discrete forecasts, or as the indicator day of a time averaged forecast, which may be the central day or the

most heavily weighted day. For instance, a typical weekly (7 day) mean would be a flat average of daily means or values over the period from $\tau - 3$ through $\tau + 3$. However, given a forecast or ensemble of forecasts produced by a numerical model out to day N , the forecast associated with validation time τ could be a weighted average in time over some or all days from 1 to N , with the weights for each day changing with τ .

In an attempt to address both aspects of seamless S2S forecasting in a single formulation. Ford et al. (2018) and Dirmeyer et al. (2018) used a time weighting based on a Poisson function:

$$P_{\tau,k} = \frac{\tau^k e^{-\tau}}{k!}, \tag{1}$$

where forecast data from multiple lead times k contributes to the averaging over multiple days of the forecast. Note that the function $P_{\tau,0}$ has nonnegligible values for small τ ; $k = 0$ corresponds to the initial state of the forecast, or more accurately the daily mean, total or extreme from the 24-h period preceding the initial condition. It is certainly admissible to include this in a forecast, representing in some sense an element of persistence from conditions at the time of the forecast, but for each τ the value of $P_{\tau,0}$ was set to 0 and the other terms renormalized so $\sum_{\tau} P_{\tau,0} = 1$ by Ford et al. (2018) and Dirmeyer et al. (2018) to focus on the performance of the forecast models. Thus, at any lead τ , from a forecast out to N days, the Poisson-weighted forecast of variable A is

$$F_{\tau} = \frac{\sum_{k=1}^N A_k P_{\tau,k}}{\sum_{k=1}^N P_{\tau,k}}. \tag{2}$$

The Poisson function applied as a weighting function in this way has some of the desirable characteristics: it is very narrow and peaked at $\tau = 1$, mimicking a discrete forecast, while transitioning to a broad averaging function approaching a Gaussian distribution for large τ . In fact, its standard deviation is simply $\sqrt{\tau}$. This gives it the property of transitioning from a front-weighted forecast at very short leads while widening and centering the averaging window as lead time increases. It applies to discrete integer values of k (days), and the integral over the entire range of k is 1 for any value of τ .

While the Poisson function is mathematically elegant, it does not have all of characteristics desired for seamless S2S forecasting. It does not have independent tunable parameters; k must be an integer, and τ must be defined in the same terms as k . Scaling τ to make $P_{\tau,k}$

wider or narrower in time introduces the need to rescale the function so the area under the curve, the sum of weights, remains equal to 1. A lognormal function could prove more tunable, but would share the problem that some averaging across days occurs even for the first day of the forecast. For most variables and applications, it is desirable to maintain discrete daily forecasts for several days, perhaps a week or longer, before transitioning to an averaging scheme. A possible solution is to delay application of the Poisson function until some arbitrary number of days into the forecast (i.e., apply an offset to τ and keep discrete forecasts for the first N days). The inflexibility of the behavior of the Poisson function could lead to such compromises.

A more graceful solution is to follow an approach like (Zhu et al. 2014) but with a tunable parameter to allow for the first several days of a forecast to be discrete daily values instead of time means. One possibility is to use a window function centered on forecast lead time τ with a width ω that grows with lead. One form of such a function is

$$\omega = 2 \text{int}(w\sqrt{\tau}) - 1. \tag{3}$$

In this case, a uniform weighting is applied over the window, such that

$$\Omega_{\tau,k} = \begin{cases} 1/\omega & 2|\tau - k| - 1 < \omega \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

One may also put a limitation on ω so that it does not widen to include forecast nonpositive leads such as the forecast initial state or past states: $\omega = \min[2 \text{int}(w\sqrt{\tau}) - 1, 2\tau - 1]$. The factor w controls the rate at which the width of the averaging window grows with lead time. In the examples below, we will examine three values: 1.0, 1.4 and 2.0.

Functions related to gamma distributions like the Poisson or lognormal functions could be advantageous where the evolution of predicted phenomena mathematically reflect those forms, such as stream discharge after intense rain. Yet the window weighting function has some advantages. Its uniform weighting across the days included in the averaging is simple and familiar. It is also tunable by choice of a single parameter w , while the Poisson function has a fixed evolution of its shape. However, the tuning is limited; this and similar single-parameter functions have the characteristic that the number of days at the start of the forecast that have no averaging applied (discrete with a window width of 1 day) has an invariant relationship to the averaging period. One may want more control over the number of days when there is no averaging applied, before the transition to a growing window at longer leads.

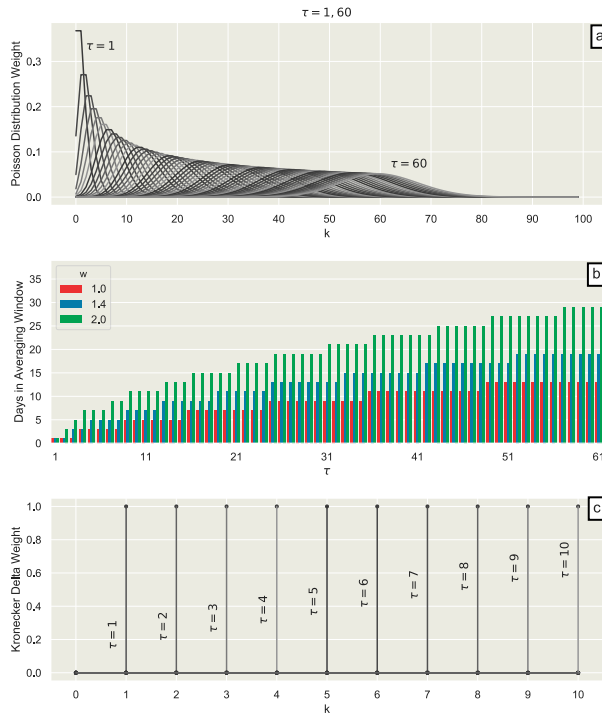


FIG. 1. (a) Poisson distribution weights $[P_{\tau,k}$ in Eq. (1)] as a function of lead time k in days for a range of forecast validation lead times τ . (b) The width of the window weighting distribution for indicated values of w as a function of τ . (c) An example of the Kronecker delta employed as a weight to represent discrete daily forecasts.

An effective weighting function exists for a discrete daily forecast—it is the Kronecker delta:

$$\delta_{\tau,k} = \begin{cases} 0 & \tau \neq k \\ 1 & \tau = k \end{cases} \quad (5)$$

It also has the properties that it applies to integers of k and its sum over all values of k is 1 for any value of τ . That means blending by any linear combination of the Kronecker and weighting functions like the Poisson or window functions, or others, as long as their coefficients also sum over time to 1, will be well behaved.

Figure 1 shows the characteristics of these functions for a selection of values of τ . Because the Poisson distribution converges to a normal distribution at large τ , we can estimate from the standard deviation that at a 30-day lead, 90% of the distribution spans the 18 days centered on day 30, at a 60-day lead the span is about $25\frac{1}{2}$ days, and at a 90-day lead, the span is a little over 31 days, showing how a Poisson weighting for forecast averaging comes to resemble a monthly mean at a one-season lead. The window function for the values of w shown spans 9–19 days at a 30-day lead and 13–29 days at a 60-day lead. Multiday averages begin after 1, 2, and 3 days for $w = 1.0, 1.4,$ and $2.0,$ respectively.

The most general approach would be to have a method to blend Kronecker and any arbitrary time-averaged distributions that accomplishes a seamless transition between discrete and time-mean forecasts that is easy to apply and flexible, so a user can customize the blending. This would be an adaptable approach for S2S applications. Such a blending function needs to be able to maintain a delta-like behavior beyond day 1, having something like an S-shape, and the inflection point should be a selectable parameter. The Hill equation (Hill 1910) satisfies these criteria, spanning the range 0–1 for all positive integers τ using only two parameters; for this application we use the form:

$$H_{\tau} = \frac{1}{\left(\frac{\tau - 1}{\alpha - 1}\right)^{\beta} + 1}, \quad (6)$$

where β controls the shape of the distribution (i.e., how quickly H_{τ} transitions from 0 to 1), and α corresponds to the forecast verification time τ at which the function crosses the central value of 0.5. Using this blending function, we have a weighting function of the form:

$$W_{\tau,k} = H_{\tau} \delta_{\tau,k} + (1 - H_{\tau}) D_{\tau,k}, \quad (7)$$

where $D_{\tau,k}$ can be $P_{\tau,k}, \Omega_{\tau,k},$ or any other distribution function whose weighting properties are conserving, and the forecast of variable A at lead τ is

$$F_{\tau} = \frac{\sum_{k=1}^N A_k W_{\tau,k}}{\sum_{k=1}^N W_{\tau,k}}. \quad (8)$$

The denominator is necessary when the lead time approaches the end of the forecast duration N , wherein the sum of the weights begins to drop significantly below 1.

Figure 2 shows several examples of distributions of H_{τ} , and Fig. 3 shows some distributions of $W_{\tau,k}$. For any value of α , a larger value of β will produce a sharper transition from discrete to time-averaged forecasts. Too large a value of β effectively reintroduces the seam into the seamless forecast. As can be seen in Fig. 3, the Hill equation can be tuned to achieve the desired combination of transition lead time and sharpness of the transition. The tuning may be chosen to maintain a specific level of uncertainty or error growth in the forecasts, a possibility not explored here, or to meet specific user needs. For example, the construction industry may want discrete forecasts out to one week to plan the deployment of materials and labor on job sites, but find the increased skill of time-averaged forecasts at longer leads useful to plan for the ordering and shipment of supplies.

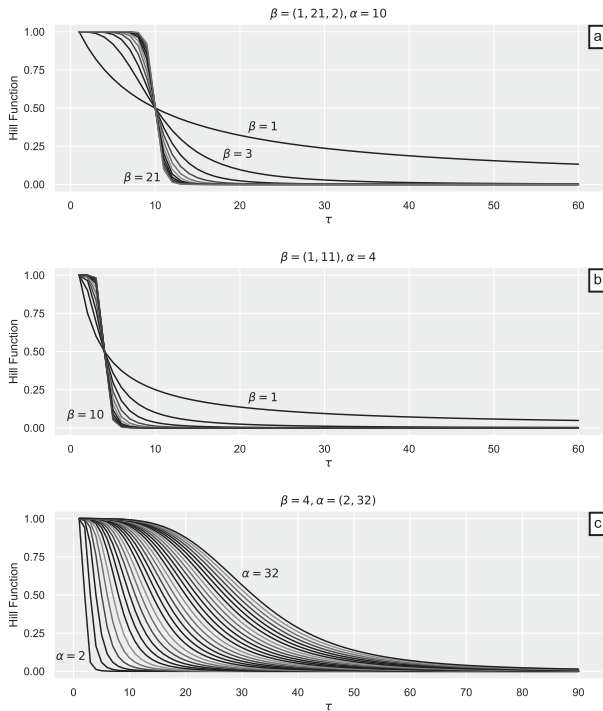


FIG. 2. Examples of the shape of the Hill function [Eq. (7)]: (a),(b) for the indicated values of α across the indicated range of β and (c) for the indicated value of β across the indicated range of α .

3. Example applications

Applications of the Poisson distribution by itself as a weighting function to create seamless predictions have been shown previously (Ford et al. 2018; Dirmeyer et al. 2018). Here we will show examples using the window function alone, the Hill function as a blending method to control the transition to time-averaged weighting, and examples involving ensemble forecast statistics that are an essential part of S2S forecasting. We do not provide an exhaustive set of seamless prediction examples, but rather a sampling to demonstrate the possibilities and the behaviors of these various approaches.

a. Considerations for validation data

One characteristic of these seamless approaches to forecast validation is that the validation data, typically either observations or reanalysis, must acquire a second time dimension. In other words, the validation states are not only a function of the calendar date, but also the lead time of forecast to be validated, because of the varying distribution functions applied. This is not as novel as it may seem. Daily versus pentad or monthly mean data has this characteristic in discrete jumps. Moving averages with different sized windows also present different values for the same calendar day. The difference here is

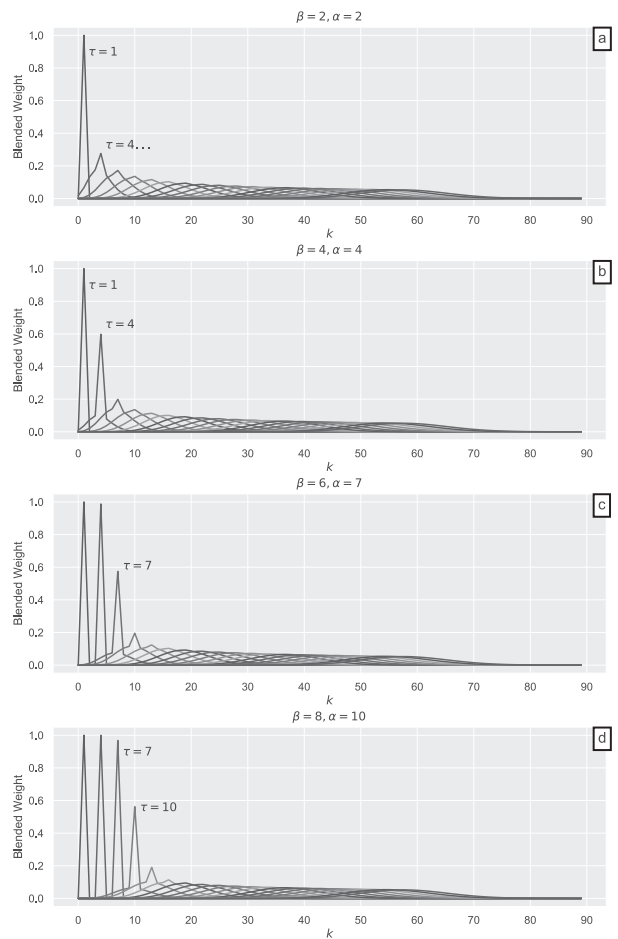


FIG. 3. Examples of the blending weight [$W_{\tau,k}$ in Eq. (8)] as a function of lead k for a range of forecast validation lead times τ (interval of 3 days starting with $\tau = 1$), with indicated choices of α and β . The examples progress from (a) a very quick transition from discrete to time averaged with lead to (d) a long period with daily forecasts transitioning to more weight toward time-averaged forecasts after a lead of 10 days.

that a full spectrum of lead times is present. If applied to the SubX model forecast data, maximum lead times range from 32 to 45 days depending on the model (Pegion et al. 2019). For the international S2S prediction project, several models archive forecasts of 60 days or more. Operationally, seasonal prediction systems may have forecast durations of several months out to a year.

Figure 4 shows how the seamless weighting using the Hill function to blend Kronecker and window distributions affects validation time series. Examples are shown for wintertime temperatures over the northern United States, and summer rainfall over central India. At long lead times relative to the value of α , individual time series smooth and the curves representing different initial conditions (ICs) converge. However, near the IC (start of each curve in Figs. 5b and 5d) the effect of the

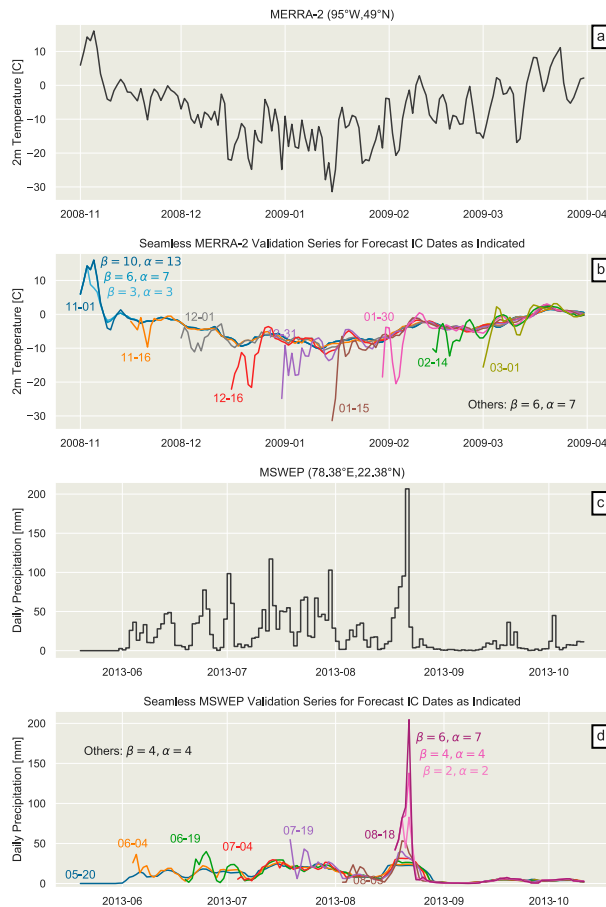


FIG. 4. Effects on sample time series of blending Kronecker and window distributions using the Hill function: (a) daily mean temperature spanning 5 months for a grid cell in MERRA-2 and (b) how the time series would appear as validation data when the weighting function $W_{\tau,k}$ is applied with values of α and β at initial times indicated by the dates color-coded with the curves. The window factor $w = 1.4$. (c),(d) As in (a) and (b), but for gridded analyzed daily precipitation from MSWEP.

Kronecker weighting and its transition to the window weighting are evident in the increased detail mirroring the daily time series. Also, for the 1 November IC of temperature and the 18 August IC for precipitation, three different choices of the Hill function parameters are shown, displaying how the delay and rate of transition can be carefully controlled. For precipitation, the ability to resolve the extreme daily rainfall total on 22 August is clearly affected by the choice of parameters. Use of other time-averaging functions like the Poisson function has similar effects (not shown).

b. Sensitivity to Hill parameters

When subseasonal predictions are validated using a range of parameters for the Hill function, we can see how skill estimates vary. Figures 5 and 6 show results for

2-m air temperature validation of the Beijing Climate Center Climate Prediction System version 1 used by the Chinese Meteorological Agency (CMA; Wu et al. 2010) from the S2S project (Vitart et al. 2017) averaged across the indicated seasons spanning 1994–2013 at grid boxes containing some major U.S. cities. Anomaly correlation coefficient (ACC) and root-mean-square error (RMSE) are the metrics shown in Figs. 5 and 6, respectively. Results pairing the Kronecker distribution with the window distribution instead of the Poisson distribution (not shown) have broadly similar characteristics. Statistics are shown across three parameters (forecast lead time, α , and β), and each of the facets shows the average across the range of the parameter perpendicular direction. Validation is for all forecasts whose validation date, defined by forecast lead time, is in the indicated season.

Common features emerge. For ACC, there is the expected decrease of skill with increasing lead time. However, the decrease in skill is clearly more gradual for low values of α (i.e., shorter lead time at which discrete to time-averaged transition occurs), and in some cases the largest ACC is at intermediate lead times of about 1 week (e.g., summer daily minimum temperature for Atlanta). Meanwhile, for large values of α , ACC decreases more rapidly with lead. Over Atlanta and some other locations over the southeastern United States (not shown), there is a minimum in ACC between 10 and 15 days lead at large α values with a gradual drop at longer leads. Atlanta also shows very poor skill at all leads for daily maximum temperature, and some apparent initialization issue affecting minimum temperature, although skill for daily means is very good in the first few days. At both very short (less than 5 days) and very long (more than 15 days) lead there is little sensitivity to the choice of α . There is also little sensitivity in skill to the choice of β at short leads, but a clear reduction in skill for low values of β (i.e., gradual transition from discrete to time-averaged) at longer leads. The bottom facet of each panel shows how skill varies across the ranges of α and β : low values of α paired with large values of β consistently result in the largest ACC values.

Similar results are seen for RMSE (Fig. 6) with slightly different emphases. The largest RMSE usually occurs at leads of around 10 days for large α values. Low values of β , especially after about 10 days, also lead to large RMSE. Otherwise features are similar as for ACC.

Recalling how α and β affect averaging, the skill characteristics are evident. Large values of β give a sharp transition from discrete to time-average forecasts. Small values of α correspond to an early transition from discrete to time-averaged forecasts. Together they act to minimize the discrete realm of the forecast evolution,

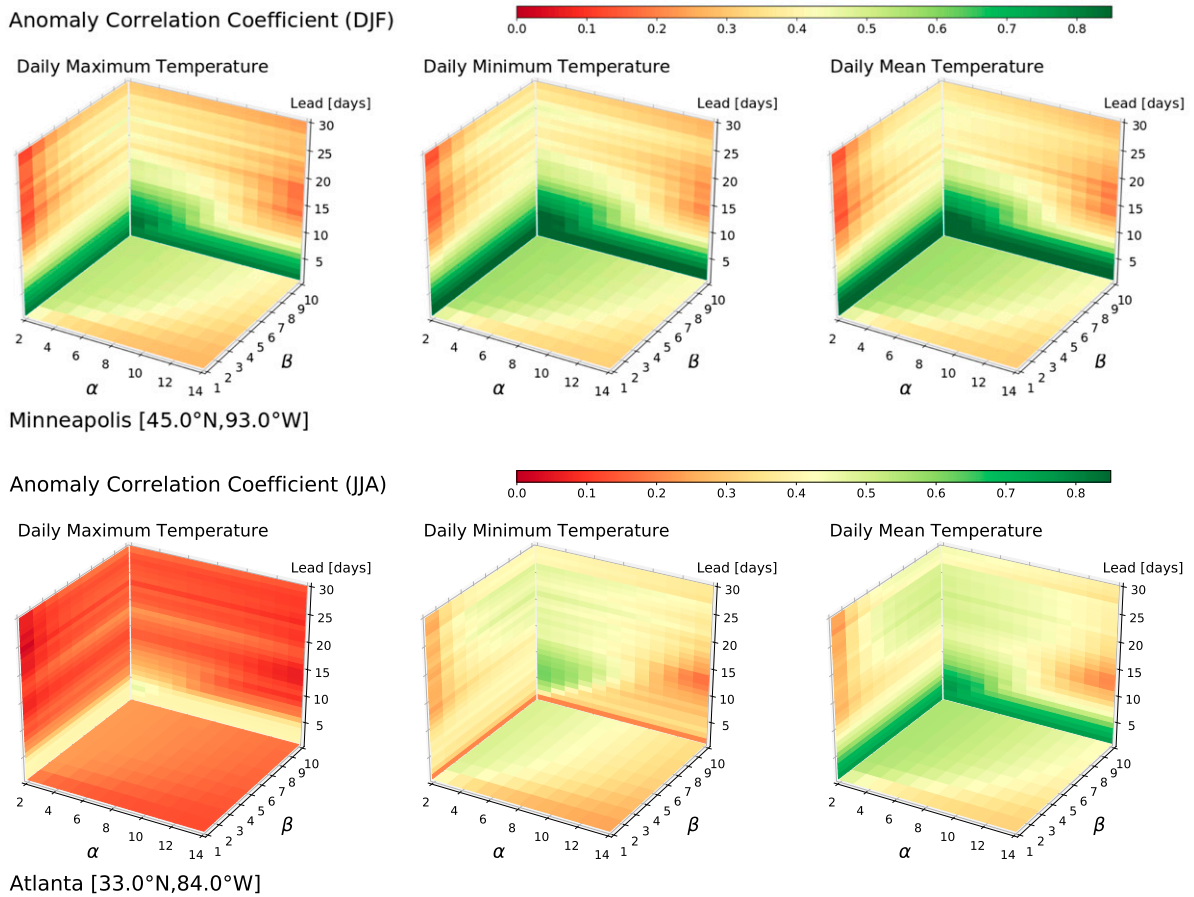


FIG. 5. Three-parameter plots of CMA forecast skill (anomaly correlation coefficient) averaged during 1999–2010 for the indicated seasons and grid cells for (left) daily maximum temperature, (center) daily minimum temperature, and (right) daily mean temperature. In each box, the color indicated the arithmetic mean along the perpendicular dimension.

creating more of a broad time-mean for verification. This favors the large-scale phenomena that affect forecasts and deemphasize event-based aspects where forecast models struggle.

However, the goal of choosing the parameters for a seamless forecast is not to boost scores, but to provide the most useful information for application of the forecasts. In Figs. 5 and 6, there is indication that model skill for discrete forecasts drops off rapidly after 5 days lead. That skill can be increased, and uncertainty decreased, with the right parameter choices, leading to time mean forecasts after just a few days, but this may not be what users want or need.

c. Ensemble forecasts

Ensemble forecast characteristics can also be employed with this method in the calculation of skill scores. In the first example, we use the NCEP Climate Forecast System (CFS) version 2 (Saha et al. 2014). CFS subseasonal

forecasts have only four ensemble members per day, so we use the discrete ranked probability skill score (RPSS_D; Weigel et al. 2007), which is unbiased by removing the effect of the nonzero expected ranked probability score found for a small random ensemble, in addition to scaling based on a climatological forecast. Forecasts for each ensemble member are constructed for Poisson weighting and various choices of α and β for the Hill function, to complement the existing discrete forecast data. We use three equal categories for temperature, based on the assumption of a normal distribution (i.e., thresholds set at $\pm 0.967\sigma$). $RPSS_D = 1$ for a perfect categorical forecast, and $RPSS_D = 0$ means no skill relative to a climatological forecast.

Additionally, we use the ECMWF system, which produces weekly initialized forecasts with 50 ensemble members as part of the S2S prediction project (Vitart et al. 2017). Heat wave occurrences are computed from deterministic ECMWF forecasts of daily maximum

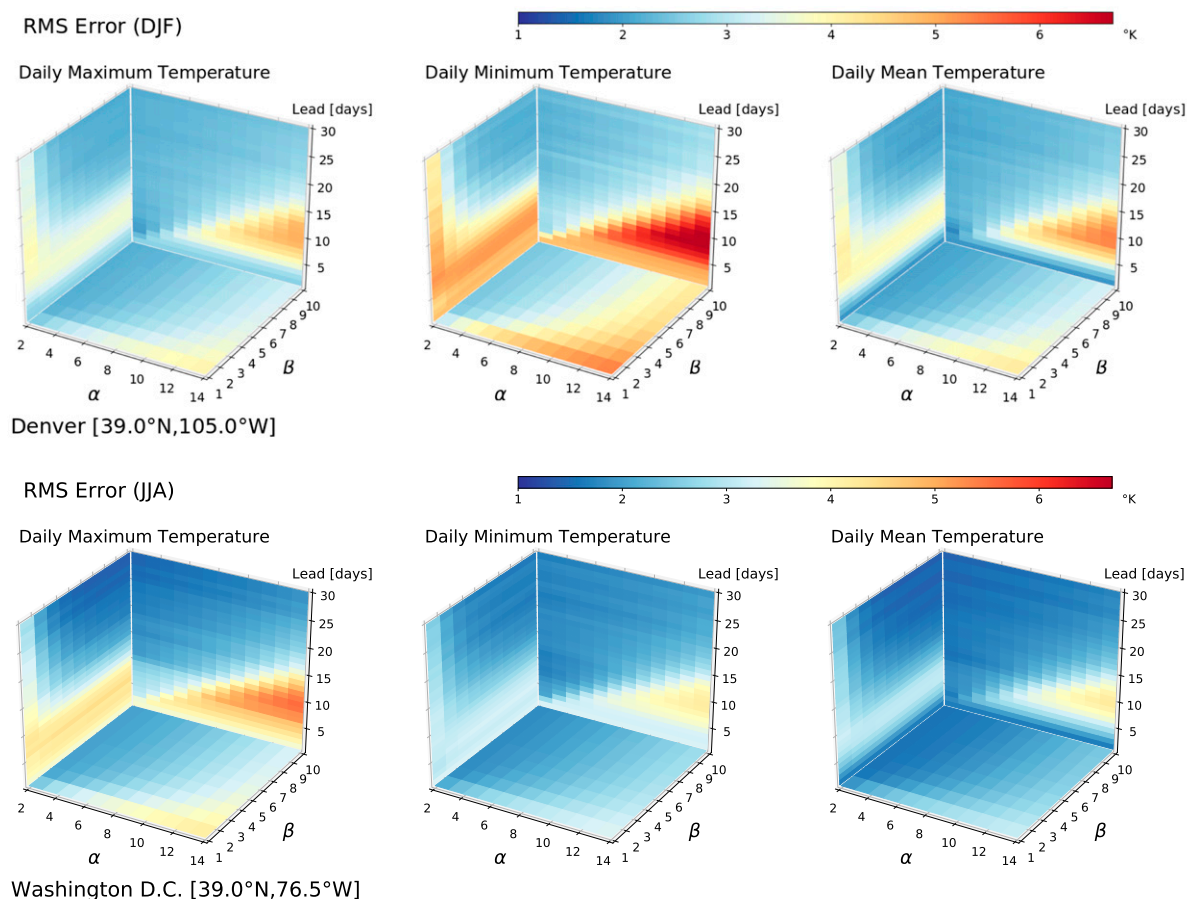


FIG. 6. As in Fig. 5, but for root-mean-square error. Also note that the locations are different than in Fig. 5.

temperature, identifying heat wave days as those which the maximum temperature exceeds the climatological 90th percentile. Maximum temperature climatologies are computed separately by ensemble member, model grid cell and forecast lead time. Similar to the treatment of NCEP CFS forecasts, ECMWF discrete, binary heat wave forecasts are constructed for each ensemble member, and an ensemble probabilistic heat wave forecast is then generated using all 50 ensemble members. To the probabilistic forecasts we apply Poisson weighting with various choices of α and β for the Hill function to demonstrate how the weighting can be used with the Brier skill score (BSS) to assess model forecast skill.

Examples for a representative grid cell are shown in Figs. 7 and 8. A number of features are found to be common for temperature and heat wave forecasts over CONUS, largely irrespective of season:

- The discrete forecast is typically competitive with weighted forecasts out to about days 4–5, but becomes clearly the worst forecast after about day 7.
- The largest spread among choices for Hill equation parameters α and β are between 7 and 14 days (Fig. 7). This illuminates week 2 as the critical period for transition. Typically, the large α values have the lowest skill—these represent maintaining the discrete forecast the longest as the primary contributor to the weighted combination.
- At long leads, small values of β often show a skill advantage, particularly for CFS temperature forecasts. These are the forecasts with the smoothest, least abrupt transition from discrete to time-averaged weighting. However, this is something of a *false skill* in that these forecasts have more weight on the very short time scales many days prior to the centered validation date. It is analogous to skill in month 1 forecasts that are derived almost entirely from the first week. Furthermore, this smooth weighting is not always an advantage—there are locations where skill is poor at very short lead times, especially for minimum temperature in this model. In these cases, as with ECMWF probabilistic heat wave forecasts,

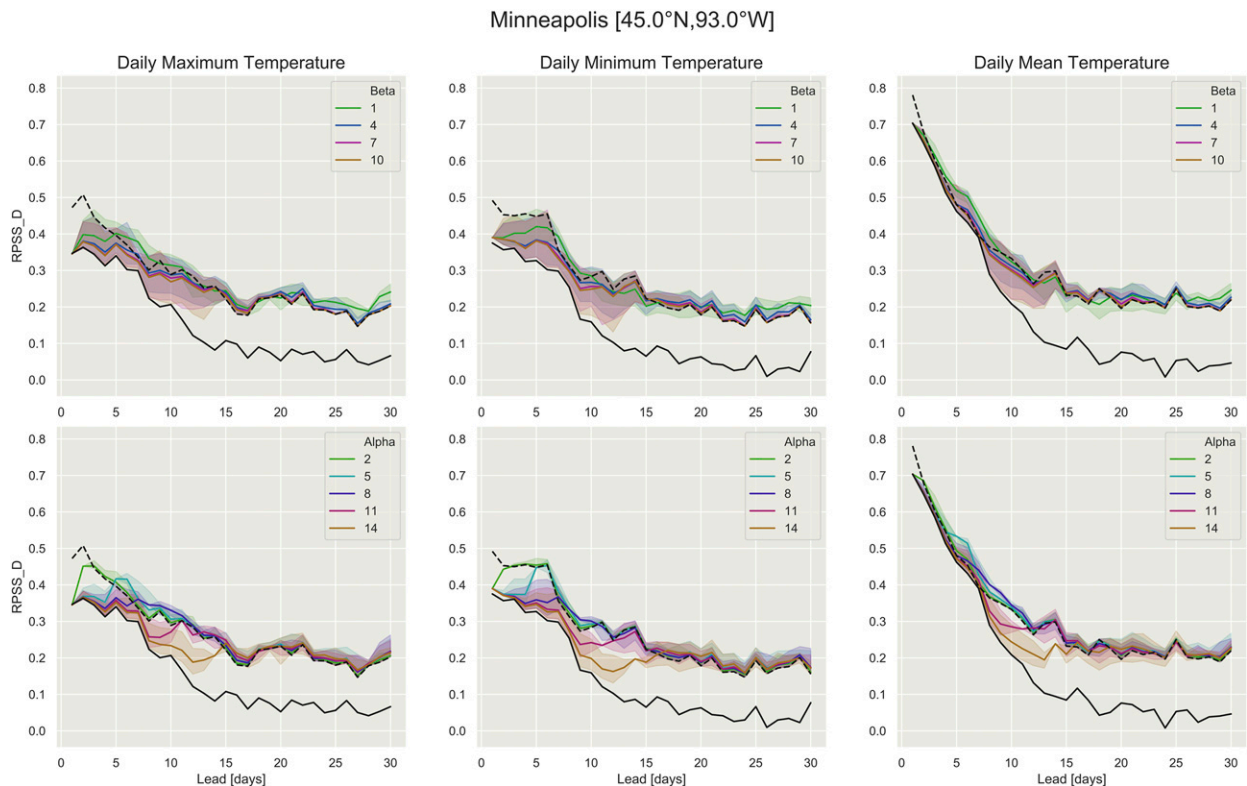


FIG. 7. Time evolution of NCEP CFS ensemble forecast skill (discrete ranked probability skill score) averaged during 1999–2010 for March–May and grid cell for the indicated temperature statistics. The solid black line is for purely discrete day-by-day forecasts, and the dashed black line is purely Poisson-weighted forecasts. (top) Mean skill scores for 4 values of β are shown with shading for the spread across 13 values of α from 2 to 14. (bottom) Mean skill scores for 5 values of α are shown with the spread across 10 values of β from 1 to 10.

the forecasts with small β values have the lowest skill scores at leads 20–30 days.

- At long leads, all choices of α and β converge to the Poisson forecast. At 25–30 days, this has not quite occurred for forecasts with some of the small β values, but by day 45 (not shown) all are indistinguishable.

In the second example, 21-member ensemble forecasts by the Environment and Climate Change Canada (ECCC) Global Environmental Model (GEM) (Lin et al. 2016) out to 30 days from the 1° resolution SubX dataset (Pegion et al. 2019) for the period of the European heat wave of 2018 are used. The window distribution by itself is applied with three different values of the parameter w are compared to discrete forecasts at a range of lead times. The ECCC-GEM forecasts are made once per week, so there is not a complete set of forecasts for all lead times validating on any particular date. The initial state (IC) is on 0000 UTC of the indicated date, so the 1-day window for 19 July initial states, for instance, is the daily mean for 19 July.

Figure 9 shows results for a single grid cell in south-central England (52°N, 1°W). Figure 9a shows the validation targets of daily mean 2-m air temperature based

on ERA5 climatology during the 40-yr period of 1979–2018, averaged to the SubX grid and as a function of the width of the averaging window in days. A heat wave is defined as a temperature at or exceeding the 95th percentile for the date, which is calculated separately for each averaging window width. Generally, as the averaging window widens, more dates are included, showing that heat wave event was persistent if not entirely consistent from day to day. Thus, many of the days may not have been in the top-two warmest for the last 40 years, but lie within longer periods that were exceptionally warm.

The remaining panels show how the effect of the forecast lead time on the width of the averaging window alters the depiction of periods of extreme heat. Figure 9b shows the number of forecast ensemble members whose discrete day-by-day temperature forecast exceeded the threshold for the given date. Figures 9c–e show similar validation statistics for different parameters of the window function; the symbol indicates the width of the window at the given lead time (refer to Fig. 1b). No bias correction has been applied to these forecasts; the intent is to demonstrate the effect of lead time and

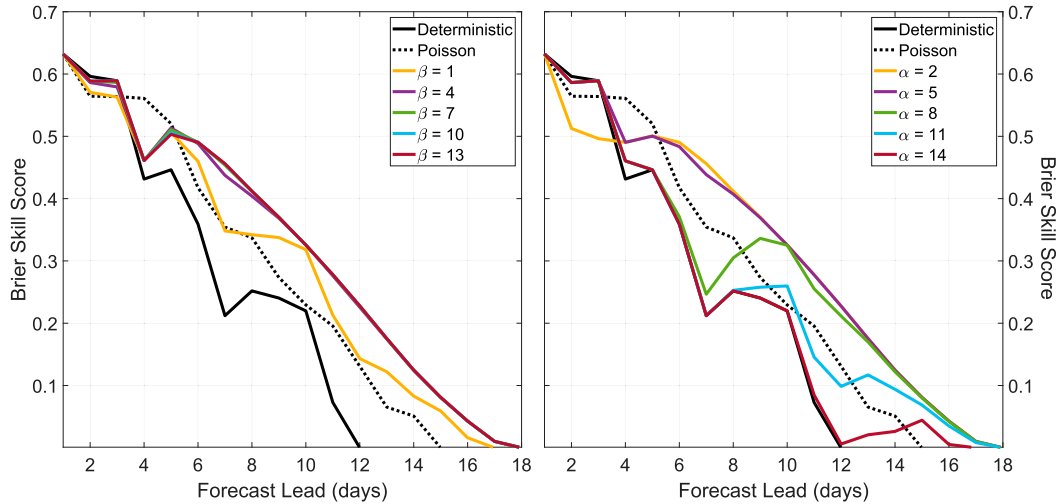


FIG. 8. Time evolution of ECMWF ensemble probabilistic heat wave forecast skill [Brier skill score (BSS)] over the period 2015–19. The solid black line is purely discrete, and the dashed black line is purely Poisson weighted. (left) BSS for 5 values of β with spread across 13 values of α from 2 to 14. (right) BSS for 5 values of α with spread across values of β from 1 to 13.

averaging period on results, not rigidly validate the model performance.

Pertinent to our original contention that precision of timing is of diminishing importance compared to capturing of the event as lead time increases, it can be seen that more ensemble members tend to predict the heat

wave at longer lead times with the widening averaging window than the discrete forecasts, which drop distinctly with each week's older initial date. This is clear both for the event centered around 7 July, and the later event centered around 25 July that contributes to the broadening of the heat wave events with wider averaging windows.

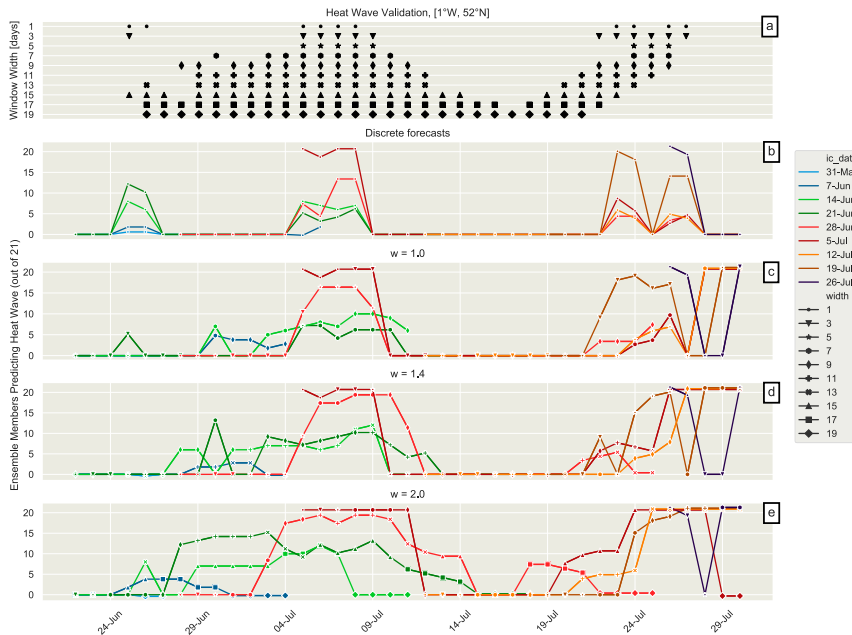


FIG. 9. (a) Symbols mark dates with a heat wave based on ERA5 (1979–2018 climatology) on the given dates in 2018 as a function of the width of the centered averaging window (each width has a unique symbol). The same symbols are used in the remaining panels where color indicates initial date of weekly ECCS ensemble forecasts: (b) discrete forecasts (no time averaging), and (c)–(e) window weighting with parameter $w = 1.0, 1.4,$ and $2.0,$ respectively. Breaks in lines occur where the window width changes.

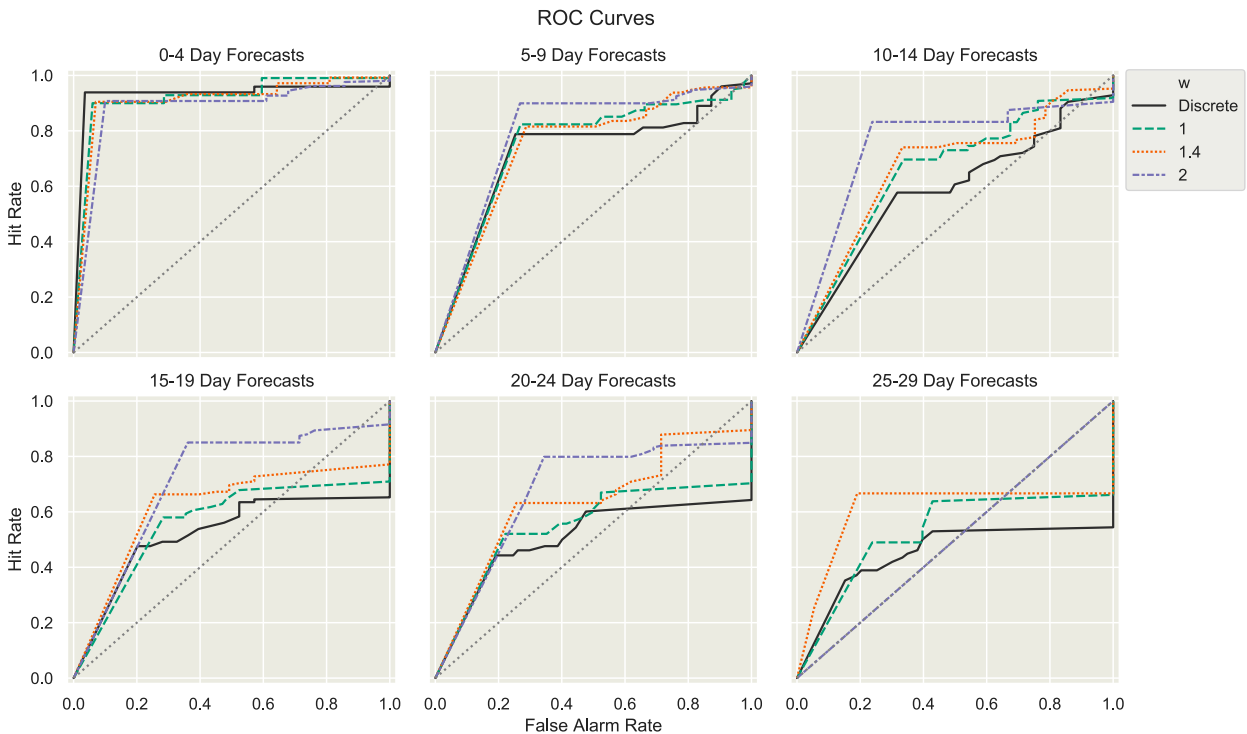


FIG. 10. Heat wave ensemble forecast statistics based on a “heat wave” vs “no heat wave” forecasts validated during 1 Jun–30 Aug 2018 for the grid cell in Fig. 9 as a function of forecast lead time and window weighting parameter w . Each panel is for all valid forecasts at the indicated lead times from ECCC.

Figure 10 quantifies these ensemble forecast statistics for this time period for the same grid cell based on heat wave (historical top 5% of temperatures for the date) versus no heat wave. Ensemble probability forecasts are validated as Relative Operating Characteristic (ROC) curves spanning a 3-month period to improve sample size. The use of the uniform weighting window that expands with time clearly improves areas under the ROC curves over unweighted discrete forecasts beyond the first 5-day validation period (more so as w increases) and increases the probability of detection. At 25–29 days lead there were no points for the $w = 2$ case besides the corners. A much larger sample would clarify these traits and is advisable before any general application of these methods as a means to select the best parameters for weighting functions.

Finally, we show an example of the 2-m temperature field and ECCC-GEM ensemble forecast statistics for the extreme heat event over Northern Europe for the validation date of 19 July 2018 in Fig. 11. Slight variations in temperature contours among panels are due to the differing number of days included in the averaging window—extreme heat thresholds are calculated separately for each window width. The rows correspond with forecast lead times of 2, 9, 16, and 23 days from top to bottom. At very short leads, nearly every grid cell has

either 21 (white) or 0 (darkest gray) members predicting extreme heat as the ensemble has not spread much; the color distribution bar at the bottom of the panels shows this dichotomy clearly. As forecast lead increases, the gray shades spread out more uniformly, but the wider averaging windows (right side of figure) tend to maintain the heat wave forecasts. White or light areas corresponding to the stippled area, where extreme heat events occur in ERA5, indicate better forecasts and better (lower) Brier scores (BS) (Brier 1950). Weighting $w = 2.0$ gives the best Brier scores; $w = 1.4$ gives slightly better scores than $w = 1.0$ out to 16 days in this case, after which $w = 1.0$ is best. The deterministic ECCC-GEM ensemble forecast at longer leads extends well south and east of the Baltic Sea (bottom-left panel), which is an area registering a heat wave in the wider averaging windows (7–17 days). This is a clear example of small errors in forecast timing at subseasonal time scales punishing the long lead forecast skill in the deterministic context, while being usefully skillful when the timing demands are relaxed.

4. Discussion

We have presented a general framework to transition numerical predictions seamlessly from discrete (valid on

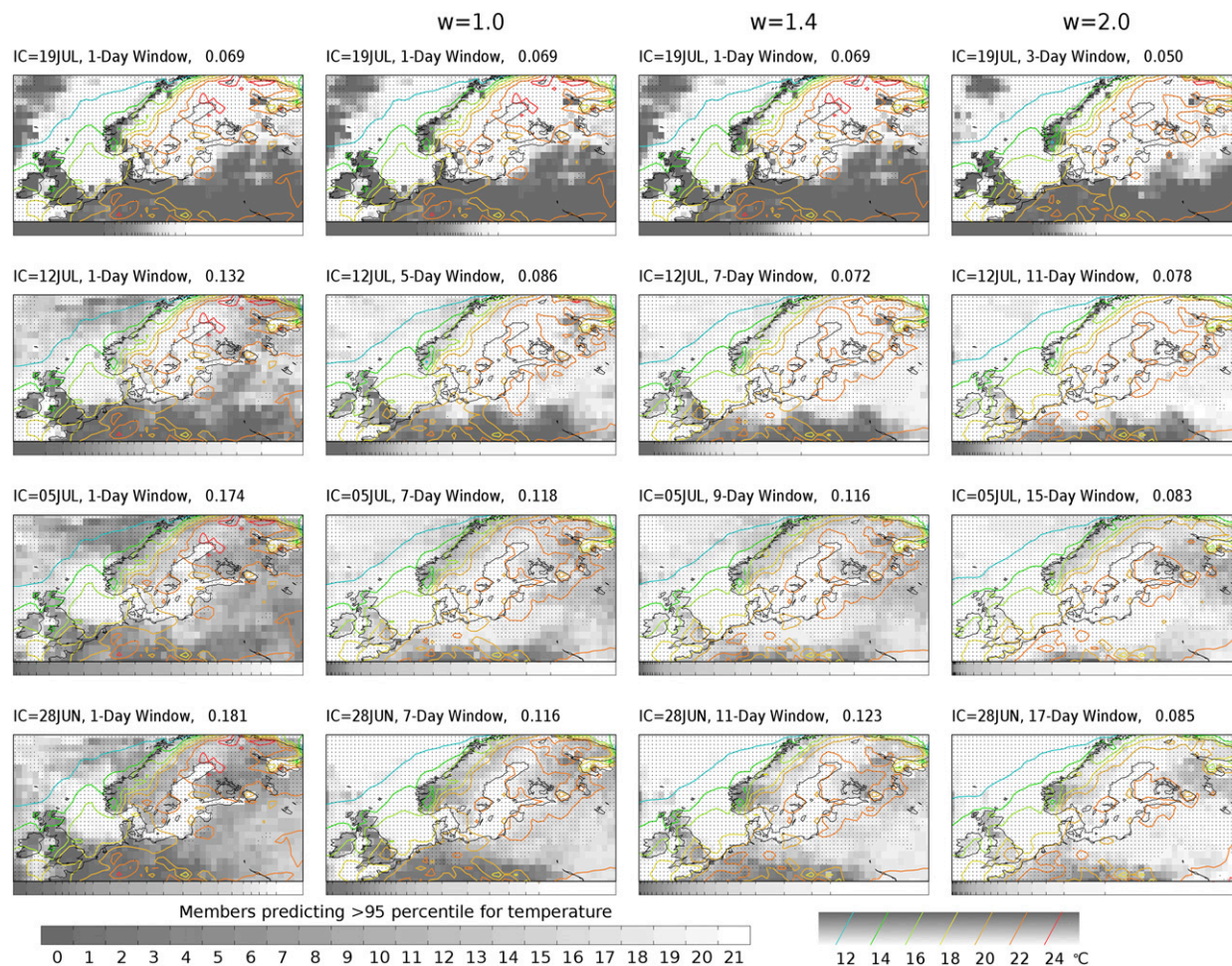


FIG. 11. Probabilistic heat wave forecast validation over Europe from the ECCC-GEM forecasts. Rows are for different initialization dates as indicated in the headers of each panel; columns are for (left) discrete forecasts of daily temperature and (remaining columns) the window distribution with the indicated w parameter yielding averages over the indicated windows centered on 20 Jul 2018. Contours show 2-m air temperature from ERA5 averaged over the same windows. Stippling shows the area where extreme heat (≥ 95 th percentile) for the averaging period exists in ERA5. Gray shades indicate the number of ensemble members predicting an extreme event for the averaging period; the color spectrum beneath each panel shows the proportional area for each value. Two Brier scores are shown above each panel, calculated over the domain.

the shortest time interval of forecast data, i.e., daily values in these examples) to time-mean (multiple days) periods of validation that can be applied to deterministic or probabilistic forecasts of both continuous and non-continuous (e.g., binary) variables on subseasonal time scales. The technique presents a general solution to transition from discrete daily weather forecasts to any sort of weighting approach that might be applied to time-average forecasts at longer leads. The two-parameter Hill function can be used to blend the two types of forecasts in a way that controls the transition rate from one to the other and the crossover point beyond which the time-averaged forecast receives more weight than the discrete forecast. This approach provides more control than either the Poisson or window weights which have been used for this

purpose in previous studies. The technique applies a Kronecker delta weight to designate discrete forecasts and examples for time-averaged forecasts using both Poisson function weights and uniform window weighting with a window that widens with increasing lead time, although any form of time-averaging can be used.

This seamless methodology does require a more complicated observed climatology, with two time dimensions, for anomaly estimation and validation. For any date, the model climatology is also a function of the forecast lead because the averaging window grows with forecast lead. This 2D time construct actually should be used for any model with time evolving errors (i.e., drift), which is to say every numerical forecast model. However, it is absolutely essential for the method described here, as the temporal

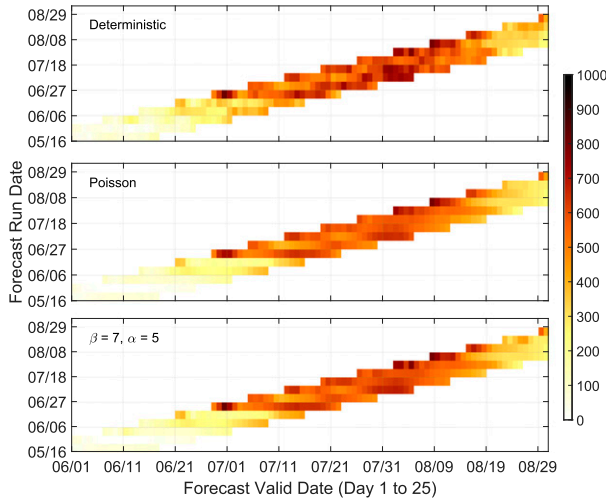


FIG. 12. Chiclet chart showing the number of CONUS grid points where GEFS maximum temperature ensemble forecasts indicate a heat wave day during summer 2012. (top) Discrete daily forecasts, (middle) purely Poisson weighted, and (bottom) weighting with values of $\beta = 7$ and $\alpha = 5$.

span of forecast validation changes greatly over lead time. With that said, the climatological behavior of model forecasts weighted using this procedure are not necessarily different from those evaluated purely discretely. This is evidenced in the chiclet chart (Fig. 12) showing forecasted maximum temperature heat wave frequency over the contiguous United States between June and August 2012 from the NCEP Global Ensemble Forecast System (GEFS; Zhou et al. 2017). However, the seamless method provides a degree of stability from forecast to forecast and between adjacent days at medium and extended ranges that discrete forecasts do not provide. GEFS maximum temperature forecast skill (ACC) verified using the seamless method persists to longer leads than when verified using either purely discrete verification or a time-invariant Poisson weighting window method, consistent with Fig. 12 (Fig. 13). Effectively, the rapid error and uncertainty growth of discrete deterministic forecasts is moderated by the seamless transition and growing averaging window (Buizza et al. 2015).

The Poisson and window functions presented here have an aspect of transition to ever widening averaging windows with increasing lead time, but on their own have limited flexibility for choosing the lead time at which discrete forecasts progress to time averages. A feature of the Hill equation is the tunability of the parameters that control the transition. One is free to choose the appropriate transition center α , and width (inversely related to β) to fit the application. The parameters should probably be different for different forecast variables (e.g., lower values of α for precipitation

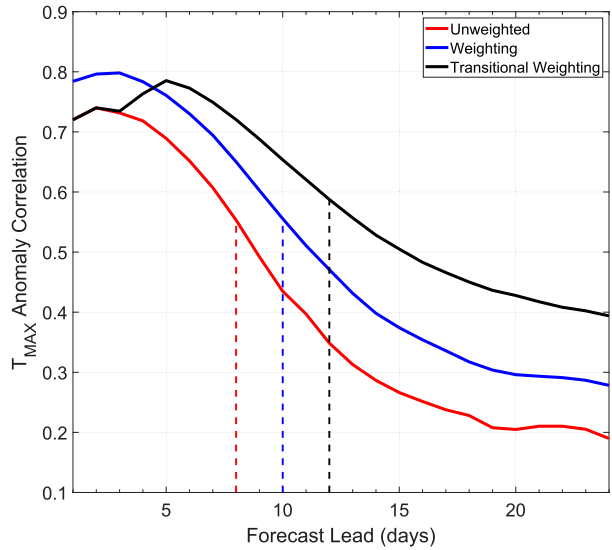


FIG. 13. GEFS maximum temperature forecast spatial anomaly correlation coefficient as a function of forecast lead time, using data from Fig. 12. Forecasts are verified using an unweighted, purely discrete verification (red line), a time-invariant, equal-weight window averaging verification (blue line), and the time-variation transitional weighting method proposed here (black line). The dashed lines indicate the lead time at which the forecast ACC falls below 0.6 based on the three verification approaches.

than temperature). They could be made to vary with season, as the range of deterministic predictability, usually the most desired form of prediction, varies throughout the year in many locations. They could also vary spatially for the output of a single forecast model, also to take advantage of differences in the range of deterministic predictability. If a uniform window averaging is used as the time-average component, an additional parameter w is available to control the evolution of the validation window with lead time. The Hill function can be applied to the Kronecker + Poisson combination, the Kronecker + window combination, or other combinations such as Kronecker + lognormal to seamlessly transition from discrete to time-averaged forecasts.

It is an open question at this time whether there is an objective, quantitative approach to optimize the choice of the parameters. The transition with lead time from discrete-dominated forecasts to time-average dominated is the “seam.” The choice of α might best be made in accordance with the point at which significant discrete forecast skill is generally lost, or to minimize the growth of forecast uncertainty with lead time. Determining such a point is a matter of estimating predictability, easier done for the forecast model than observations. Predictability loss in models could potentially be used as an emergent constraint (Hall et al. 2019) to estimate the transition point in observations.

Meanwhile, the sharpness of the seam is controlled by the value of β . Its choice may be more of a subjective one. What this approach to seamless transitioning accomplishes is to level the increase in uncertainty with forecast lead—if uncertainty can be quantified, perhaps most practically quantified by forecast error or ensemble spread, it could be used as a metric to choose β . Regardless of the choices, at very long leads the Poisson functional form dominates, and converges to a normal distribution, center-weighted on the validation day. The window function likewise spreads with lead time. This makes a smooth time-mean quantity with a window that widens gradually with lead time, as shown in Fig. 1. Using the Hill function to transition from discrete (Kronecker delta) to time-averaged (window function) forecasts provides three separate parameters and thus 3 degrees of freedom to tailor the seamless approach to the subseasonal forecast situation.

One could imagine the best choice of both parameters would depend on the application. The balance most amenable for emergency responders might be different than for the power generation industry, or the general public. On the other hand, it means forecasts can be tailored to multiple applications, as the Hill function provides great versatility with only two parameters. All that is needed is the two sets of forecasts in hand: the discrete day-by-day model output, and the Poisson or window distribution average. The final form of the seamless forecast is simply a linear combination of the two, and different versions of a final forecast can be produced from the same two sets as a postprocessing step. Whether the proposed approach is advantageous in any particular situation would need to be explored on a case by case basis, just like any other existing model methods. The ability to compensate smoothly for the natural increase in uncertainty with forecast lead could be useful in many applications. The generalized technique presented here is not meant to construct forecasts with the highest skill, but to construct forecasts with the highest utility across time scales from weather to subseasonal in a single seamless product.

Acknowledgments. This research was supported by the NOAA Climate Program Office from the Modeling, Analysis, Predictions, and Projections (MAPP) Program Grants NA16OAR4310095 and NA16OAR4310066. Data from the WWRP/THORPEX-WCRP joint research project on Subseasonal to Seasonal (S2S) prediction were taken from the archive maintained at ECMWF (<https://confluence.ecmwf.int/display/S2S>). Data for the Subseasonal Experiment (SubX; <http://cola.gmu.edu/kepegion/subx/index.html>) were taken from the archive at

the International Research Institute for Climate and Society.

REFERENCES

- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Brunet, G., and Coauthors, 2010: Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **91**, 1397–1406, <https://doi.org/10.1175/2010BAMS3013.1>.
- Buizza, R., M. Leutbecher, and A. Thorpe, 2015: Living with the butterfly effect: A seamless view of predictability. *ECMWF Newsletter*, No. 145, ECMWF, Reading, United Kingdom, 18–23, <https://www.ecmwf.int/sites/default/files/elibrary/2015/17265-living-butterfly-effect-seamless-view-predictability.pdf>.
- Dirmeyer, P. A., S. Halder, and R. Bombardi, 2018: On the harvest of predictability from land states in a global forecast model. *J. Geophys. Res. Atmos.*, **123**, 13 111–13 127, <https://doi.org/10.1029/2018JD029103>.
- Ebert, E., and Coauthors, 2013: Progress and challenges in forecast verification. *Meteor. Appl.*, **20**, 130–139, <https://doi.org/10.1002/met.1392>.
- Ford, T. W., P. A. Dirmeyer, and D. O. Benson, 2018: Evaluation of heat wave forecasts seamlessly across subseasonal timescales. *npj Climate Atmos. Sci.*, **1**, 20, <https://doi.org/10.1038/s41612-018-0027-7>.
- Hall, A., P. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nat. Climate Change*, **9**, 269–278, <https://doi.org/10.1038/s41558-019-0436-6>.
- Hill, A. V., 1910: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.*, **40**, iv–vii.
- Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, **139**, 573–584, <https://doi.org/10.1002/qj.1991>.
- Lin, H., N. Gagnon, S. Beaugard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian meteorological centre. *Mon. Wea. Rev.*, **144**, 4867–4883, <https://doi.org/10.1175/MWR-D-16-0138.1>.
- Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate Atmos. Sci.*, **1**, 4, <https://doi.org/10.1038/s41612-018-0014-z>.
- Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003: A study of subseasonal predictability. *Mon. Wea. Rev.*, **131**, 1715–1732, <https://doi.org/10.1175/2558.1>.
- Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, <https://doi.org/10.1175/BAMS-D-18-0270.1>.
- Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, <https://doi.org/10.1175/JCLI-D-12-00823.1>.
- Shapiro, M., and Coauthors, 2010: An earth-system prediction initiative for the twenty-first century. *Bull. Amer. Meteor. Soc.*, **91**, 1377–1388, <https://doi.org/10.1175/2010BAMS2944.1>.
- Shukla, J., R. Hagedorn, M. Miller, T. N. Palmer, B. Hoskins, J. Kinter, J. Marotzke, and J. Slingo, 2009: Strategies:

- Revolution in climate prediction is both necessary and possible: A declaration at the World Modelling Summit for climate prediction. *Bull. Amer. Meteor. Soc.*, **90**, 175–178, <https://doi.org/10.1175/2008BAMS2759.1>.
- Vitart, F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, **140**, 1889–1899, <https://doi.org/10.1002/qj.2256>.
- , and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- WCRP, 2005: The World Climate Research Programme strategic framework 2005-15: Coordinated Observation and Prediction of the Earth System (COPES). WCRP-123, WMO/TD-1291, World Meteorological Organization, 65 pp., https://www.wcrp-climate.org/documents/WCRP_stratempl_LowRes.pdf.
- Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, <https://doi.org/10.1175/MWR3280.1>.
- Wu, T., and Coauthors, 2010: The Beijing Climate Center atmospheric general circulation model: Description and its performance for the present-day climate. *Climate Dyn.*, **34**, 123–147, <https://doi.org/10.1007/s00382-008-0487-2>.
- Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP global ensemble forecast system in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, <https://doi.org/10.1175/WAF-D-17-0023.1>.
- Zhu, H., M. C. Wheeler, A. H. Sobel, and D. Hudson, 2014: Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.*, **142**, 1556–1569, <https://doi.org/10.1175/MWR-D-13-00222.1>.