Multimodel Subseasonal Forecasts of Spring Cold Spells: Potential Value for the Hazelnut Agribusiness

STEFANO MATERIA,[a] ÁNGEL G. MUÑOZ,[b] M. CARMEN ÁLVAREZ-CASTRO,[a] SIMON J. MASON,[b] FREDERIC VITART,[c] AND SILVIO GUALDI[a,d]

[a] *CSP Division, Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*
[b] *International Research Institute for Climate and Society, and the Earth Institute at Columbia University, New York, New York*
[c] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*
[d] *Istituto Nazionale di Geofisica e Vulcanologia, Bologna, Italy*

(Manuscript received 26 April 2019, in final form 23 October 2019)

ABSTRACT

Producing probabilistic subseasonal forecasts of extreme events up to six weeks in advance is crucial for many economic sectors. In agribusiness, this time scale is particularly critical because it allows for mitigation strategies to be adopted for counteracting weather hazards and taking advantage of opportunities. For example, spring frosts are detrimental for many nut trees, resulting in dramatic losses at harvest time. To explore subseasonal forecast quality in boreal spring, identified as one of the most sensitive times of the year by agribusiness end users, we build a multisystem ensemble using four models involved in the Subseasonal to Seasonal Prediction project (S2S). Two-meter temperature forecasts are used to analyze cold spell predictions in the coastal Black Sea region, an area that is a global leader in the production of hazelnuts. When analyzed at the global scale, the multisystem ensemble probabilistic forecasts for near-surface temperature are better than climatological values for several regions, especially the tropics, even many weeks in advance; however, in the coastal Black Sea, skill is low after the second forecast week. When cold spells are predicted instead of near-surface temperatures, skill improves for the region, and the forecasts prove to contain potentially useful information to stakeholders willing to put mitigation plans into effect. Using a cost–loss model approach for the first time in this context, we show that there is added value of having such a forecast system instead of a business-as-usual strategy, not only for predictions released 1–2 weeks ahead of the extreme event, but also at longer lead times.

1. Introduction

Subseasonal forecasts are extended-range weather forecasts: a few times a week, general circulation models (GCMs) used for short-range forecasts are extended for 30–45 days. In recent years, this time scale ranging between the limit of deterministic predictability (which is usually set to 10–14 days; Lorenz 1982) and the season (i.e., 60–90 days) has been given particular attention. Indeed, such forecasts can provide crucial advance warning to decision-makers about forthcoming weather events (Batté et al. 2018), while application-ready capabilities could allow many sectors (e.g., energy, transport, agriculture) the opportunity to systematically plan on a new time horizon (White et al. 2017).

The weather time scale is considered a pure atmospheric initial-condition problem, while the seasonal to interannual range depends strongly on the slowly evolving components of the Earth system, such as ocean temperatures. Subseasonal variability fills the gap between the two, and it has always been considered a challenging time range for predictions, since the lead time is sufficiently long to dilute the information imparted by the atmospheric initial conditions, and it is too short for the memory of the ocean to influence the climate system (Vitart et al. 2017). However, potential sources of predictability for this time range have been identified, mostly the Madden–Julian oscillation [MJO; e.g., Vitart and Molteni (2010) and references therein], the state of

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

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-19-0086.s1.

Corresponding author: Stefano Materia, stefano.materia@cmcc.it

DOI: 10.1175/WAF-D-19-0086.1

El Niño–Southern Oscillation (ENSO, e.g., Liang and Lin 2018) and their interconnection (Hoell et al. 2014), as well as soil moisture (Guo et al. 2011; Koster et al. 2011), snow cover (Thomas et al. 2016; Orsolini et al. 2013), sea ice (Furtado et al. 2016), stratosphere–troposphere interactions (Tripathi et al. 2015), and cross-time-scale interference of multiple climate drivers (Muñoz et al. 2015, 2016, 2017). Despite the increasing knowledge on sources of predictability, and its socioeconomic importance, subseasonal forecasting is still at a relatively early stage of development. Climate models' prediction skill at such time scales is in fact still modest at mid- and high latitudes (DeFlorio et al. 2018).

Nonetheless, the experience acquired through years of seasonal forecast research and operations show that increasing the ensemble size allows for a wider sampling of the possible weather/climate evolution (Palmer et al. 2000), enhancing the probabilistic forecast skill (Buizza 2008). Due to various constraints, very few institutes can run 45-day forecasts several times a month with a large ensemble set. Efforts like the Subseasonal to Seasonal Prediction project (S2S) database (Vitart et al. 2017), which freely provides a set of subseasonal forecasts and hindcasts produced by 11 different prediction systems, are helping the scientific community to advance understanding of sources of predictability, model improvement and forecast skill. Many studies have demonstrated the enhanced forecast quality of multimodel ensembles compared to a more conventional single-model ensemble approach (Hagedorn et al. 2006; DelSole and Tippett 2014; Vigaud et al. 2017). In addition, a large sample size allows the use of consensus between the different model forecasts to get some insight into the predictability (Piedelievre 2000), provides for an insightful evaluation of probabilistic skill (Krishnamurti et al. 2006), and imparts a potential economic value to the forecast (Richardson 2000; Alessandri et al. 2011).

Although a few studies have recently evaluated the subseasonal forecast quality and potentialities in a multisystem setting, both for precipitation (Vigaud et al. 2017) and for temperature (Ferrone et al. 2017), there is still no assessment regarding cold extremes at such time scale.

After the first two weeks, the aggregation of outputs into weekly values tends to increase the skill of most atmospheric field predictions (Rodwell and Doblas-Reyes 2006). The weekly time frame is often sufficient to detect a cold spell: although many cold events last two or three days, especially those occurring early in boreal spring, their impact on weekly temperature anomalies is identifiable. The 2–3 day time frame is shorter than the typical time span established by the Expert Team on Climate Change Detection and Indices (ETCCDI) for cold spell duration (i.e., six consecutive days of minimum daily temperatures lower than the 10th percentile; Alexander et al. 2006), but a few hours below freezing are enough to heavily harm crops and fruit farms (Rodrigo 2000).

Hazelnuts, for instance, are vulnerable to severe frost in late winter and the beginning of spring, when female flowers have just begun their development, and low temperatures may be destructive for germination (Ustaoglu 2012; Beyhan and Odabas 1996). In particular, in 2004 and 2014 two abrupt extreme cold spells hit the coastal Black Sea at the end of March, causing profound damages to hazelnut plantations there. Curbing chances for plants to fructify, more than half of the annual harvest was lost and hazelnut prices increased sharply (Erdogan and Aygün 2017). If farming of one commodity is concentrated in a small region, such localized and ephemeral weather events can easily affect its worldwide production. Hazelnut agribusiness is in fact highly centralized in the southern and eastern coastal Black Sea, where about 70% of the world's production is found [Food and Agriculture Organization of the United Nations (FAOSTAT) 2016]. This high vulnerability, strongly affecting prices' volatility, urges an exploration of the opportunity to predict such extreme events through targeted subseasonal forecasts, which are able to provide early notice of possible weather hazards.

The main motivation of this study is the implementation and assessment of a methodology to supply the agribusiness sector in general, and the one in the coastal Black Sea as a case study, with timely and reliable information at subseasonal time scales. Since the nature of this method requires a large ensemble set of subseasonal forecasts, the main objectives of this study are

- to explore a consistent multimodel approach for temperature-related variables in the context of S2S, and evaluate it through forecast verification metrics;
- to assess the quality and potential economic value of cold spell seasonal forecasts during the most critical season for nuts production.

2. Methodology

a. Multimodel and observational data

In this study, we make use of several models participating in the S2S database (Vitart et al. 2017), to ensure a robust number of ensemble members. The selection of the models participating in the multisystem ensemble was based on two criteria:

1) models should have reforecast periods that overlap for the longest number of years, so that robust statistics can be obtained;

TABLE 1. Characteristics of the seasonal forecast systems participating in the multisystem. The last row corresponds to the multisystem.

Model	Institution	Country	Ensemble size	Start dates
BCC-CPS-S2Sv1	CMA	China	9	27–28 Feb, 1 Mar 13–15 Mar 30–31 Mar, 1 Apr
GloSea5	UKMO	United Kingdom	7	1 Mar 17 Mar 1 Apr
IFS Cy43r3	ECMWF	Europe	10	28 Feb 14 Mar 28 Mar
CNRM-CM 6.0	Météo France	France	14	1 Mar 15 Mar 1 Apr
Multisystem (MSys)			40	1 Mar 15 Mar 1 Apr

2) models should have at least 45-day-long integrations to allow forecast verification on periods longer than a month.

Based on these conditions, four models—with a total of 40 realizations—were selected for the multisystem (MSys; see the appendix for a list of abbreviations and acronyms used in this paper) for the analysis of cold spell events; Table 1 summarizes the ensemble size of each GCM and forecast start dates. The horizontal resolution of the dataset is a 1.5° regular latitude–longitude grid, and the reference reforecast length is 19 years, 1996–2014. We selected three start dates at the beginning of the meteorological boreal spring (1 March, 15 March, and 1 April), with each integration spanning a time lapse of 42 days (i.e., six aggregated weeks). To consider a robust sample size when assessing skill, since the forecasts are assumed to be independent, the three initializations were concatenated; the resulting time series are equivalent to 57 (19 years for the three start dates) 42-day-long forecasts per model.

The reference data for comparison and evaluation of the forecast quality are provided by the ECMWF interim reanalysis (ERA-Interim, hereinafter ERAI; Dee et al. 2011), which will be referred to as ''observations'' below. The spatial resolution of the dataset is approximately 80 km (T255 spectral), but it has been reduced to about 150 km to match with the forecast data.

b. *Near-surface temperature and definition of the cold spell index*

A simple analysis of temperature anomalies predicted in the area of interest is insufficient for an evaluation of cold spells. Nonetheless, as a reference, we evaluated the MSys near-surface (2 m) temperature skill, both at the global scale and for a region including the coastal Black Sea.

At the regional scale, we derived a cold spell index based on the study of Peings et al. (2013), customized for the region with the vastest hazelnut-farmed lands in the world. This region is located on the coastal Black Sea, and we refer to it as the northern coast of Turkey (NCT). The Turkish provinces where hazelnuts are cultivated, together with their model grid representation, are shown in Fig. 1.

To define the cold spell index, we consider a geographical domain approximately corresponding to these provinces (see boxes in Fig. 2). Since the area identified by this domain is too small for a reliable interpretation of subseasonal forecast outcomes, we enlarged its extent by selecting nearby grid points that shared similar climate anomalies during the reforecast period (1996–2014). Hence, we calculate temperature anomaly correlations (TAC) using the Pearson coefficient between each of the boxes and the surrounding points over the reference period. In this way we obtain maps of correlations, where regions characterized by higher values share similar thermal features with NCT (Fig. 2).

We arbitrarily retain regions whose TAC is higher than 0.85, in order to implement further calculations over areas large enough to be suitable for the type of analyses performed. The robust covariance guarantees that the climate variability of the enlarged region is strongly associated with that of NCT. This operation is repeated for each of the 40 ensemble members and each of the six weeks of the subseasonal hindcasts (starting on 1 March, 15 March, and 1 April), as well as for the observations. We thus obtain 40 correlation maps per start date, plus one for the observations, for every forecast week.

When more than 20% of the grid points featured by $TAC_{gp} > 0.85$ have temperature anomalies colder than

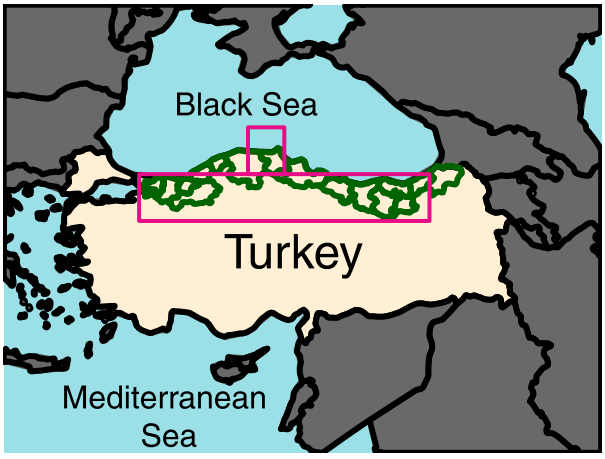

FIG. 1. Turkish provinces (in green) where hazelnuts are farmed for commercial purposes. In purple, the box used for most the analyses carried out in this work.

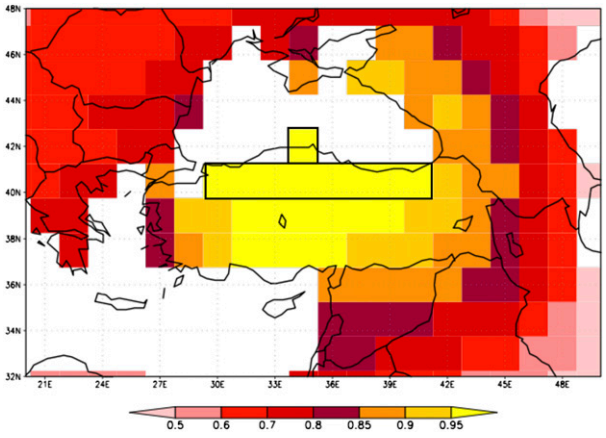

FIG. 2. Example of 2-m temperature correlation maps for a random ensemble member and a random start date. Shadings indicate correlation with the nut farms' grid points, averaged in the black box. CSPI was calculated over the area points with $r > 0.85$ (orange and yellow shades in this map).

the 10th percentile (Peings et al. 2013), calculated over the observed (for ERAI) or forecast (for MSys) temperature distribution of the grid point itself, a cold spell is detected.

The cold spell index defined above is characterized by the following magnitude and extent:

- The magnitude (MGN) of the cold spell is determined by the average temperature anomaly of the pixels with temperature $<$ 10th percentile;
- The extent (EXT) of the cold spell is the fraction of area with temperature $<$ 10th percentile (always greater than 0.2 by construction).

We can define a cold spell power index (CSPI) by

$$\mathrm{CSPI} = \mathrm{MGN} \times \mathrm{EXT}. \tag{1}$$

Figure 3 illustrates the steps to obtain the CSPI definition in a simple schematic. The choice of using the 10th percentile threshold to identify the occurrence of cold spells is not meant to diagnose temperatures dropping below zero or truly affecting plant phenological cycles. Varying thresholds to account for such local features are beyond the scope of this work and may be considered in studies that use higher horizontal resolution.

c. *Metrics to assess forecast quality and value*

1) SCORES FOR THE MULTISYSTEM EVALUATION

To assess the quality of probabilistic subseasonal forecasts, we focus on measuring attributes that any good prediction should have reliability, resolution, uncertainty, and discrimination.

Ignorance (IGN), an information theory–based verification metric, is selected because it simultaneously measures reliability (REL), resolution (RES), and uncertainty [UNC; see Weijs et al. (2010) for an extensive dissertation]:

$$\mathrm{IGN} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}. \tag{2}$$

Reliability is a measure of the conditional bias in the forecast probabilities and is 0 for a perfectly calibrated forecast. Ideally, the observed frequency equals the forecast probability for all of the issued forecast probabilities:

$$\mathrm{REL} = \frac{1}{N}\sum_{k=1}^{K} n_k D(\bar{o}_k \| f_k). \tag{3}$$

Here, the difference between the observed frequency distribution $\bar{o}_k$ and the forecast probability mass distribution f_k, both in the category k, is expressed in terms of relative entropy D, also known as Kullback–Leibler divergence (Kullback and Leibler 1951); N is the total number of forecasts issued, K is the number of unique forecasts issued, and n_k is the number of forecasts with the same probability category.

Resolution measures the amount of uncertainty in the observation explained by the forecast. The minimum resolution is 0, which occurs when the climatological probability is always the forecast or the forecasts are completely random; practically, it can be seen as the amount of information in the forecast:

$$\mathrm{RES} = \frac{1}{N}\sum_{k=1}^{K} n_k D(\bar{o}_k \| \bar{o}), \tag{4}$$

where relative entropy D is calculated between the conditional and marginal probabilities of occurrence.

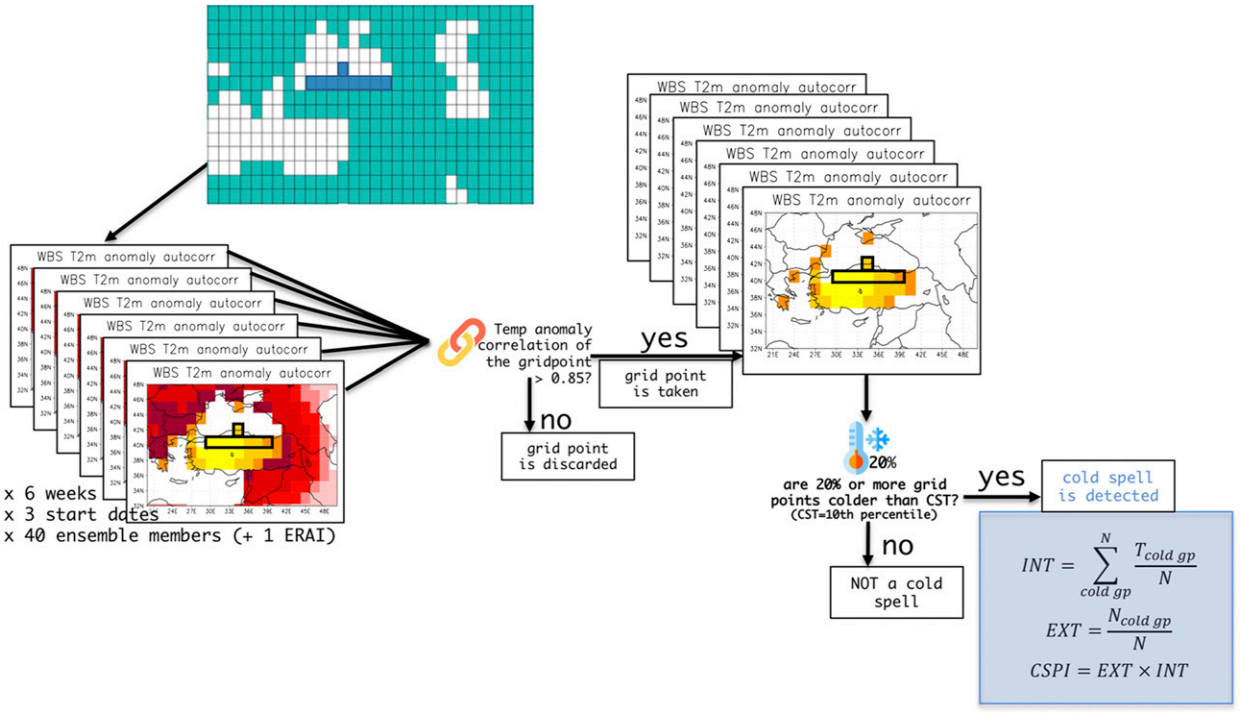

FIG. 3. Schematic of the methodology used to define a cold spell. For each start date, week, and ensemble member a correlation map is computed; in every map, only grid points with a correlation higher than 0.85 are kept for cold spell determination. If no less than 20% of these grid points have temperatures < CST (cold spell threshold, i.e., 2-m temperature is in the 10th percentile), that member is characterized by a cold spell, defined in the bottom right box (icons made by Freepik and Vectors Market from www.flaticon.com).

Uncertainty measures the initial observational uncertainty about the event through the entropy of the climatological distribution $H(o)$. Being a function of the observational climatology, it does not depend on the forecast; the uncertainty is maximum if the probability of occurrence is 0.5 and 0 if the probability is either 0 or 1:

$$\mathrm{UNC} = H(o) = -\sum_{i=1}^{n} (\overline{o}_i) \log(\overline{o}_i), \tag{5}$$

where n is the number of categories in the probabilistic forecast ($n = 3$ in this study).

Due to its relationship to Shannon's information entropy, IGN is frequently used as a proxy for forecast utility, or the amount of information gain expected from a forecast (Roulston and Smith 2002). Due to its easy interpretation, the ignorance skill score (ISS) is used here:

$$\mathrm{ISS} = -\frac{\log_2 p_k}{\log_2 n}, \tag{6}$$

where p_k denotes the probability of the realized category. This definition of the ignorance skill score is negatively oriented; locations where ISS > 1 contain less information than the climatology (ISS $= 1$), and locations with ISS < 1 contain more information than climatology. A perfect forecast has zero IGN and ISS.

The generalized relative operating characteristics (GROC) is used to assess discrimination of tercile-based probabilistic forecasts. GROC is a particular case of the two-alternatives forced choice score (2AFC; Mason and Weigel 2009), and measures "the proportion of all available pairs of observations of differing category whose probability forecasts are discriminated in the correct direction" (Mason and Weigel 2009).

These metrics are computed using the International Research Institute for Climate and Society (IRI) Climate Predictability Tool (CPT; Mason and Tippett 2019), once the probabilistic forecasts have been computed by simple counting.

In addition, a deterministic evaluation is carried out using a time correlation, computed according to the Spearman coefficient operating on the ensemble mean of each subseasonal forecast (i.e., each start date) and the corresponding observations at every grid point (Wilks 2011).

2) SCORE OF THE CONTINGENCY TABLES

To assess the MSys forecast performance, we need to verify whether the predicted cold spells were in fact

recorded by ERAI reanalysis. If the forecasts were deterministic, a retrospective prediction would either identify or not a cold spell defined by CSPI. However, because the forecasts are probabilistic, an additional choice is required: how high must the forecast probability be to trigger a cold spell alarm? In other words, do we consider a forecast of a cold spell one for which, for instance, CSPI has 16% chance of occurrence? To assess the skill of CSPI forecasts we make use of the Gerrity skill score (GSS; Gerrity 1992; Gandin and Murphy 1992), a categorical score that measures the quality of a given index at capturing the values in each of the multiple categories (Siebert 2016). The first step for the assessment of GSS consists in the assignment of a value to each of the four possible outcomes between the forecasts and observations:

1) A cold spell is predicted by the MSys hindcast and is then verified by ERAI [hit event (HE)].
2) A cold spell takes place in a specific week, but the MSys was not able to predict it [missed event (ME)].
3) The MSys predicts a cold spell, which is not recorded in ERAI [false alarm (FA)].
4) CSPI is neither predicted nor occurs [correct rejection (CR)].

These four outcomes will populate a so-called contingency table, a 2 × 2 table where the main diagonal (from top left to bottom right) contains HE and CR, in other words cases in which the hindcast and reanalysis agree, while the antidiagonal (from bottom left to top right) contains FA and ME, namely, wrong forecast assessments.

GSS provides the following scores for the elements of the contingency table:

$$e_{\mathrm{HE}} = \left(\frac{1-p_w}{p_w}\right), \tag{7}$$

$$e_{\mathrm{CR}} = \frac{1}{(e_{\mathrm{HE}})}, \quad \text{and} \tag{8}$$

$$e_{\mathrm{ME}} = e_{\mathrm{FA}} = -1, \tag{9}$$

where p_w is the 10th percentile threshold (i.e., 0.1) needed to define the CSPI.

3) Cost–loss model

To examine the potential economic benefit of the subseasonal forecasts, we make use of a simple cost–loss model (Richardson 2000). We consider a decision-maker sensitive to spring cold spells: if the cold spell occurs, the decision-maker loses part of the harvest, incurring a loss L. However, the decision-maker may decide to take action against the cold spell (a farmer could use antifrost turbines, while a buyer could purchase hazelnuts before the frost comes); in this case, there will be a cost C to take action, but L will be avoided.

Supposing the fraction of CSPI per week σ is known for past seasons: having no additional information the decision-maker could choose to apply a mitigation plan every week, with an average expenditure $E_{\mathrm{always}} = C$, or never take action, and the average loss would be $E_{\mathrm{never}} = \sigma L$. Hence, the best strategy would be

$$E_{\mathrm{no-info}} = \min(C, \sigma L). \tag{10}$$

Of course, the best case scenario would be a perfect knowledge of future weather, where the action is put in place only when CSPI will take place:

$$E_{\mathrm{perfect}} = \sigma C. \tag{11}$$

Relying on subseasonal forecasts, the decision-maker will not be able to nullify the costs but could get closer to E_{perfect}, minimizing the expenses and losses. Here the contingency table comes in to help (Table 4): the mean expense of using the forecast is obtained by multiplying HE, FA, ME, and CR by their correspondent expenditure:

$$E_{\mathrm{forecast}} = \mathrm{HE} \times C + \mathrm{FA} \times C + \mathrm{ME} \times L + \mathrm{CR} \times 0. \tag{12}$$

The price difference between E_{forecast} and $E_{\mathrm{no\text{-}info}}$ is a measure of the value of the forecast for the decision-maker. Relative to the perfect scenario, where the action is put in place only when the cold spell occurs, the value V of a forecast is

$$V = \frac{E_{\mathrm{no\text{-}info}} - E_{\mathrm{forecast}}}{E_{\mathrm{no\text{-}info}} - E_{\mathrm{perfect}}}, \tag{13}$$

which can be expanded making use of Eqs. (5)–(7) in

$$V = \frac{\min(C, \sigma L) - \mathrm{FR}(C/L)(1-\sigma) + \mathrm{HR}\sigma(1-C/L) - \sigma}{\min(C, \sigma L) - \sigma(C/L)}, \tag{14}$$

where the false alarm rate is FR = FA/(FA + CR), and the hit rate is HR = HE/(HE + ME), namely, the number of false alarms (hit events) over the total number of nonoccurred events (occurred events, i.e., CSPI).

3. Results

a. Validation of the multisystem

The models' global 2-m temperature deterministic skill in target weeks 2 and 5 is shown in Fig. 4. As explained in section 2, we use 57 seasons, instead of only

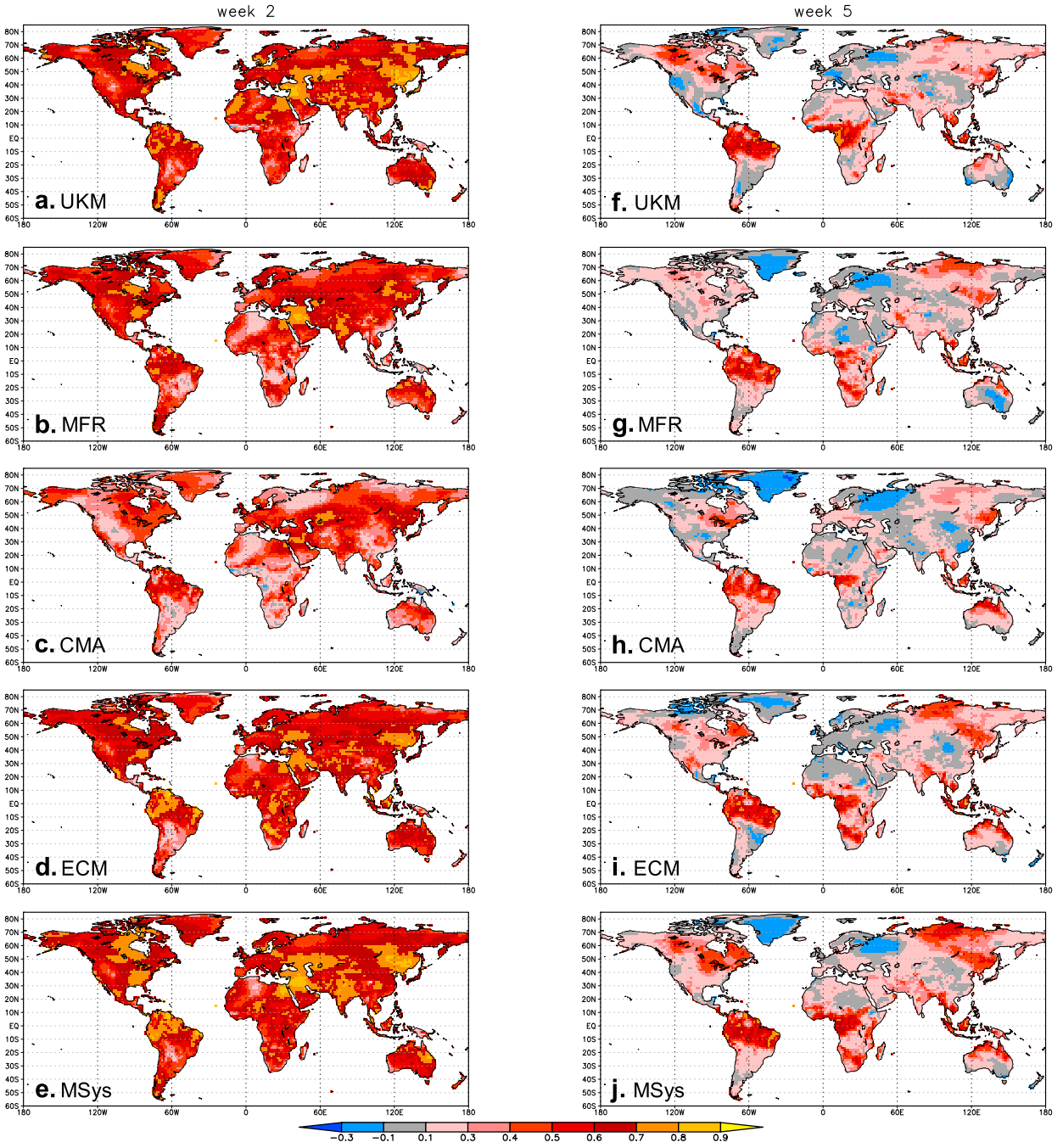

FIG. 4. Forecast skill (Spearman correlation) for (a)–(d),(f)–(i) global 2-m temperatures predicted by each single model and (e),(j) the multisystem ensemble. A concatenation of the three start dates (1 Mar, 15 Mar, and 1 Apr) was used to calculate correlations to enlarge the sample to 57 years. Shown in (a)–(e) is skill at week 2, and (f)–(j) show skill at week 5.

19, to increase the sample size available for skill assessment, an approach that has been recently used by Muñoz et al. (2018) for similar purposes.

At week 2, model skill is at a maximum in the Northern Hemisphere's low and midlatitudes, particularly in central Asia, the Middle East, northeast China, India, and eastern North America. The only equatorial area showing homogeneously high correlations is the western Amazon basin, while the rest of the tropics display fluctuating skill among different models. Models hardly agree with reanalysis in central South America and Indonesia, as well as in many parts of Africa where

the lack of data, though, could affect the reanalysis itself. CMA shows lower skill than the other models, with major weaknesses in North America, northwestern Russia and Scandinavia, Australia, and southeastern Asia (Fig. 4c). As expected, correlations are noticeably lower at week 5 in all prediction systems, and skill discordance among models reduces. However, rain forests, southeastern Asia, and northeastern China maintain significant forecast quality in all models at longer lead times, showing that the system skill—as measured by Spearman correlation—can go well beyond the deterministic limit when outputs are aggregated over a weekly time range.

The 40-member MSys shows an overall improvement of the subseasonal forecast skill with respect to each of the single models, particularly for longer-range predictions. At week 2, most midlatitude continental lands, southern Africa, South Asia including Indonesia, eastern Australia, and great part of the Amazon basin show correlations higher than 0.7 (Fig. 4e), while the large majority of the remaining land shows values larger than 0.6. The whole subequatorial band, southern Africa, the Great Lakes region, and northeastern Asia preserve significant forecast skill at week 5 (Fig. 4j).

A frequent way to provide uncertainty information in forecasts is to use a probabilistic format (Doblas-Reyes et al. 2000). As indicated before, we use the ignorance skill score to measure reliability, resolution, and uncertainty of the probabilistic near-surface temperature predictions of the MSys. Regions showing low ignorance (i.e., good skill; see blue shades in Fig. 5) during the first weeks of the forecasts tend to show decreasing skill with lead time, with several tropical locations still skillful at week 6, consistent with previous studies (Li and Robertson 2015). Nonetheless, some regions exhibiting worse-than-climatology ISS values (red shades in Fig. 5) in week 1 tend to show climatological values toward week 6, that is, an increase in skill. This is due to the fact that the models tend to be overconfident during the first weeks of the forecasts, and then converge toward climatological values (in white in Fig. 5). Hence, except perhaps at the global scale, no generalization on the tendency of probabilistic skill for spring near-surface temperature should be made when referring to any particular region. In the case of the entire Black Sea basin, a closer analysis (not shown) indicates that near-surface temperature probabilistic skill quickly degrades with lead time, suggesting that the MSys temperature forecasts are only useful during approximately the first two weeks of prediction. As shown in the next subsection, it is possible to extract actionable information from these forecasts when a different but related variable is used.

The forecasts' discrimination, as assessed by GROC, is better than climatological values (GROC > 50%, red shades in Fig. 6) for most of the tropical Americas, the Maritime Continent, and Africa, even in week 6. Since discrimination (and resolution) is considerably higher in the tropics, the evolution of GROC in Fig. 6, together with the ISS analysis described above, indicates that the forecasts require more calibration to obtain functional discrimination in the extratropics, where the region of interest is located. The role of calibration in the type of subseasonal forecasts used in this research is out of the scope of this study and will be addressed elsewhere.

It is also interesting to explore how the multisystem performs in the location where the hazelnuts are found. Figure 7 shows the model deterministic skill in the domain that could potentially host a grid point participating in the CSPI calculation (see section 2b), that is, the domain in Fig. 2. Here, time correlation averaged over the domain, for all of the six weeks of each of the start dates (Figs. 7a,b,c) is shown for every model and the multisystem, allowing us to visualize the forecast quality per target week. Increasing the sample size to 57 seasons (Fig. 7d) allows for a reduction of random data fluctuations. Correlation drops dramatically after the deterministic limit of skill (Lorenz 1982), with correlations normally not exceeding 0.4 beyond the week-2 lead time. This is in line with other studies carried out within the European domain. An attempt to reforecast the July 2015 heat wave (Ardilouze et al. 2017) concluded that the prediction system involved could not guarantee a skillful forecast more than 12–14 days before the beginning of the episode. Besides, Monhart et al. (2018) found out that, in Europe, spring is the season characterized by the worst subseasonal prediction skill.

In general, correlations decrease with time, but differences in skill are hardly appreciable between weeks 3 and 5, while a steep downward step is noticeable at week 6. The MSys is always among the three best models, often the best performing in the first two forecast weeks.

b. *Prediction of cold spells in the coastal Black Sea*

Table 2 shows CSPIs in the NCT enlarged regions for the six March and April lead weeks corresponding to the 15 March start date (similar tables for 1 March and 1 April can be found in Tables S4 and S5 in the online supplemental material). Forty-six cold spells are detected in ERAI during the 14 analyzed weeks (while there are 18 forecast lead weeks, that is, six per start date; only 14 are measurable in observations, since four weeks overlap in the 1 March and 15 March start dates). In some years, spring is not affected by any cold spell,

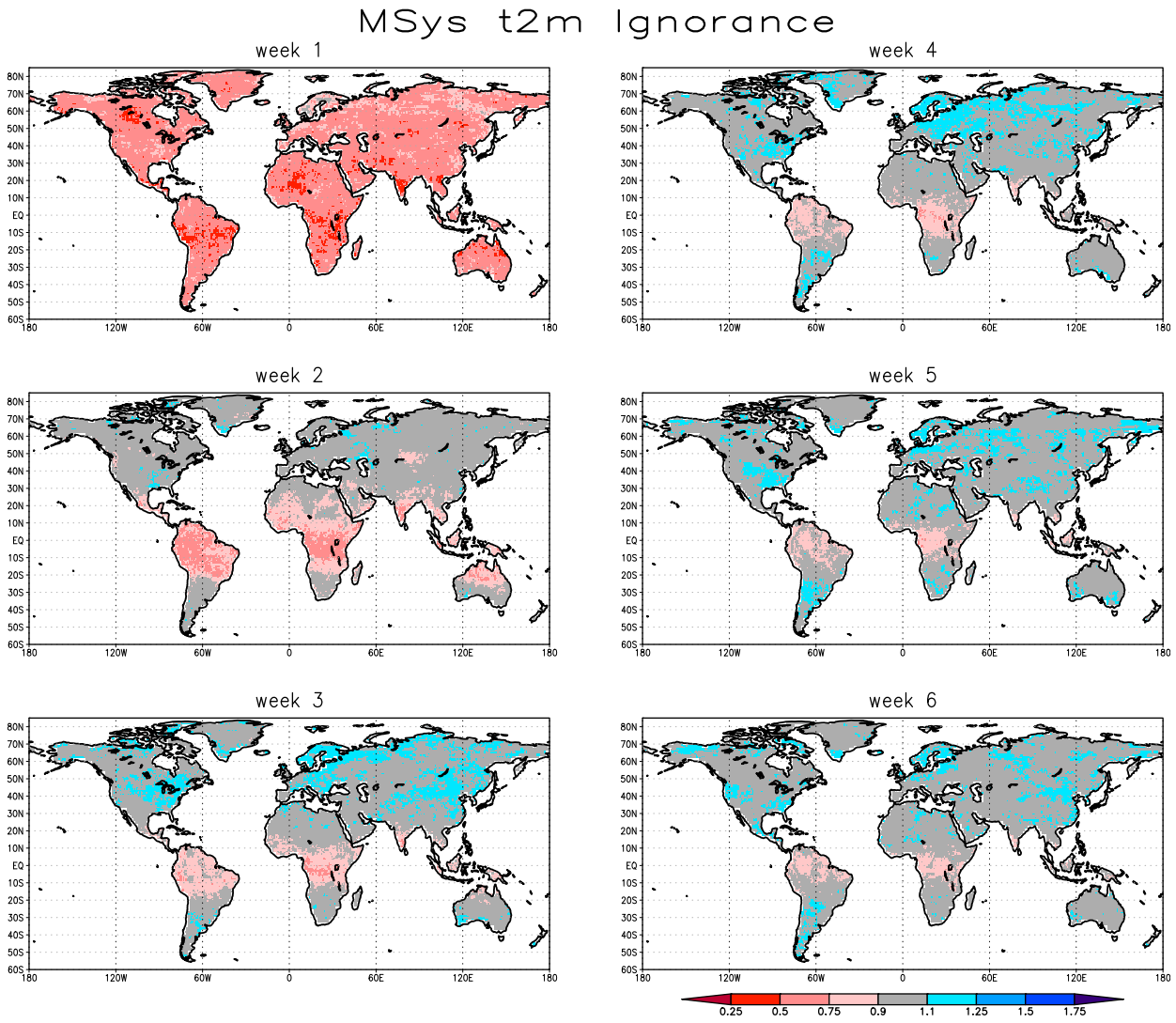

FIG. 5. Multisystem ignorance skill score for near-surface temperature during boreal spring. Blue (red) regions exhibit better (worse) skill than climatological values, which is shown in white; a perfect forecast has a value of zero ignorance skill score.

and cold spells become rarer after 2005, especially in the early part of the season. However, most of the late events (May, see Table S5) take place from 2005 onward: two events are tracked in the first nine years, and four since 2005.

Table 3 shows the probability associated with CSPIs retrospectively predicted on the start date of 15 March (Tables S4 and S5 show the same for the start dates of 1 March and 1 April, respectively). The probabilistic forecast provides the percentage of ensemble members forecasting a cold spell (a method commonly known as ''simple counting,'' including one additional ensemble member that is split between the various outcomes based on their climatological probabilities), as defined by Eq. (1), in the six following weeks. There is an evident disparity in forecast confidence between the deterministic forecast time (weeks 1–2) and the other weeks.

For the analysis, we consider 114 forecasts for each start date, that is, six weeks over the 19-yr reference period. On week 1 and week 2 the system is often overly confident. This means that in the first forecast weeks, cold spell events are predicted by most of the ensemble members, then the probability of occurrence given by the dynamical system is rather high (often above 40%). Similarly, a prediction of no cold spell is often shared, in the first forecast weeks, by the quasi totality of the members, resulting in a 0% (or very close to 0) probability of occurrence.

The spread between members amplifies in the following weeks as they drift apart from the initialization date. Only once during weeks 5 and 6 does more than a

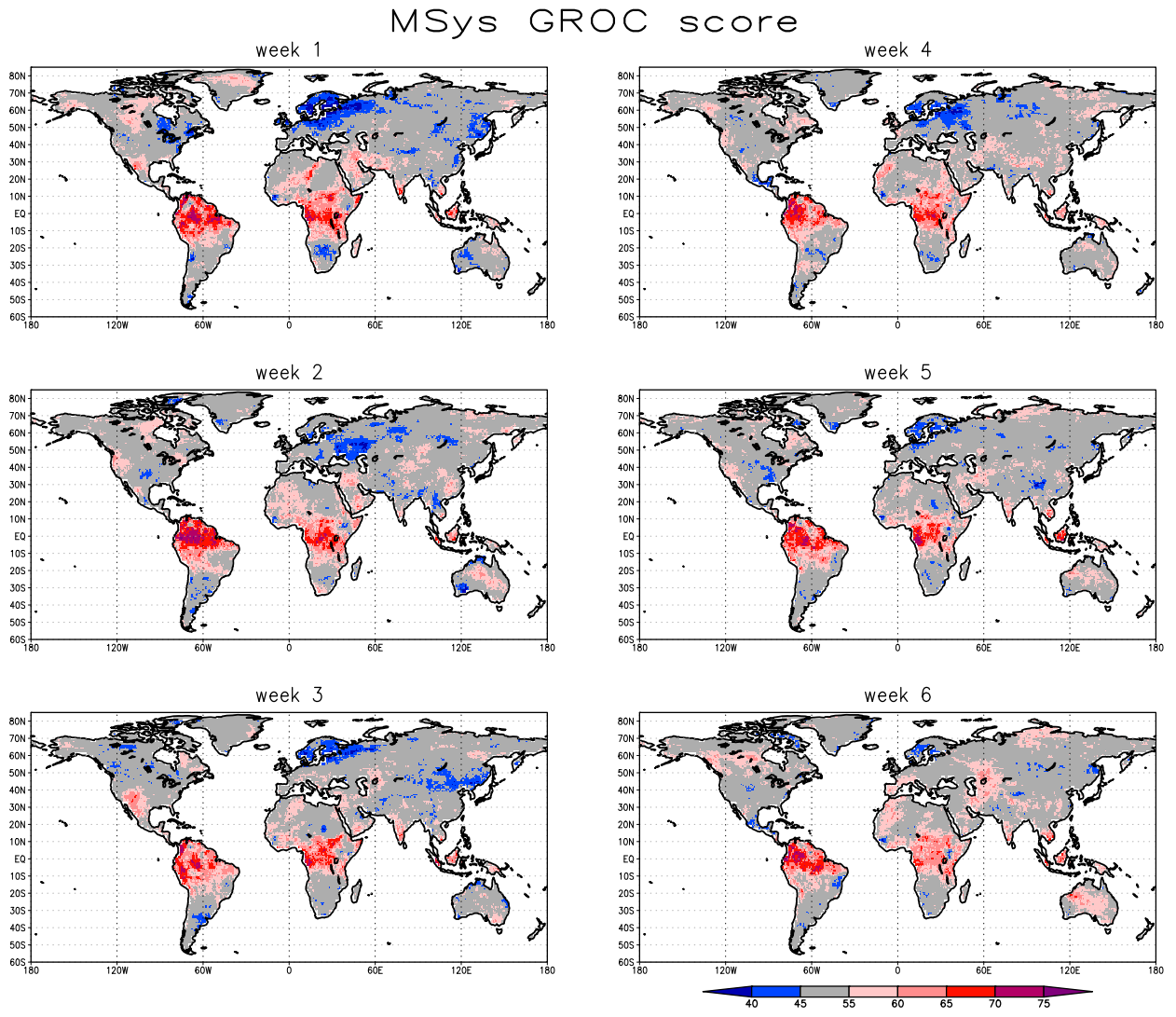

FIG. 6. As in Fig. 5, but for GROC. Values above 50 (in blue) indicate better discrimination than climatology (white), and vice versa (in red).

third of the ensemble set agree on the occurrence of a cold spell (in the 1 March 1996 forecast, Table S4). Similarly, only in very limited cases do all members agree on zero CSPI chance after week 2, and essentially never after week 4. This behavior is expected in a system subjected to a random perturbation of the initial conditions, where the error shows an amplification that increases in time (Lorenz 1963).

To take into consideration the decrease in forecasts' confidence with time, that is, the increase in the ensemble members' dispersion, we treated weeks 1–2, weeks 3–4, and weeks 5–6 separately. The sum of the 38 GSSs, relative to the 19 outcomes (1996–2014) for each two-week chunk, is maximized by changing the threshold triggering a forecast of CSPI. Consequently, for each start date we obtain the three lowest probability limits (LPLs), each associated with one two-week chunk, which set the minimum forecast probability required for a cold spell alert. In other words, a CSPI forecast probability > LPL can be considered a deterministic forecast of a cold spell.

The aggregated contingency table (Table 4) shows the realization of each of the four outcomes (HE, FA, ME, CR) separated by lead time. When the cold spell event occurs in the reanalysis (HE + ME), the system is able to predict it 48% of the time. More generally, the MSys forecast agrees with reanalysis over 77% of the time (HE + CR = 263 out of 342), with HE events being about 12% of the CR events.

The MSys produces HE + FA = 77 predicted cold spells in the NCT region, with correct forecasts recorded about 36% of the time. Considering the predicted

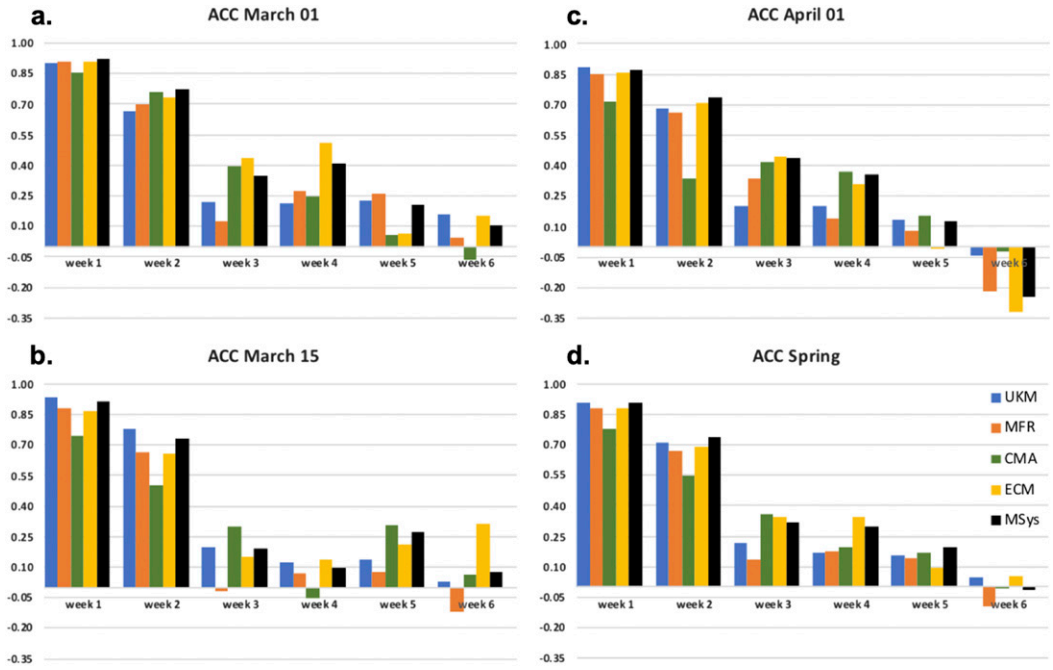

FIG. 7. Forecast skills (Spearman correlation to the ERAI reference) for 2-m temperatures predicted by each single model and the MSys ensemble averaged the region around Turkey (20.0°–49.5°E, 32.0°–48.0°N; see Fig. 2 for the exact domain location): (a) 1 Mar start date, (b) 15 Mar start date, (c) 1 Apr start date, and (d) concatenation of the three start dates.

nonevents (ME + CR), the forecast is correct almost 90% of the time.

When a cold spell is detected in the reanalysis, the MSys provides a correct forecast about 48% of the time [HE/(HE + ME)]. When CSPI does not occur (FA + CR =284), the system foresees its occurrence 49 times (total number of false alarms), that is about 17% of the observed nonevents. The forecast probability of having a cold spell is considerably higher when cold spells occur, meaning the forecast is different depending on the outcome; hence, the forecast has good discrimination.

TABLE 2. Cold spells identified by the CSPI index in ERA Interim: values for NCT for the target weeks of the 15 Mar start date.

	ERA Interim CSPI in the northern coast of Turkey (°C)					
	16–22 Mar	23–29 Mar	30 Mar–5 Apr	6–12 Apr	13–19 Apr	20–26 Apr
1996	—	—	—	−1.2	−1.9	−3.1
1997	−1.1	−4.8	—	−6.5	−2.4	—
1998	−3.3	—	—	—	—	—
1999	—	—	—	—	—	—
2000	−0.9	—	—	—	—	—
2001	—	—	—	—	—	—
2002	—	—	—	−1.0	—	—
2003	−2.2	−4.6	—	—	—	−1.0
2004	—	—	−4.8	—	—	—
2005	—	—	−4.4	—	—	—
2006	—	—	—	—	—	—
2007	—	—	—	—	−3.8	−2.1
2008	—	—	—	—	—	—
2009	—	—	—	—	—	—
2010	—	—	—	—	—	—
2011	—	—	—	−0.8	—	−0.7
2012	—	—	—	—	—	—
2013	—	—	—	—	—	—
2014	—	—	—	—	—	—

TABLE 3. Forecast probability of having a cold spell, in all the reforecast weeks of the 15 Mar start date. The lowest probability limits (LPLs) required to call for a cold spell forecast are shown under the indication of the two regions. Values in bold font indicate a hit, values in italic a false alarm, values in bold italic font a missed event, and values in normal font a correct rejection.

	Probability of predicted CSPI (15 Mar)					
	LPL = 18.6%		LPL = 13.7%		LPL = 10.8%	
	16–22 Mar	23–29 Mar	30 Mar–5 Apr	6–12 Apr	13–19 Apr	20–26 Apr
1996	0%	17.5%	*27.5%*	**20%**	**15%**	**20%**
1997	***12.5%***	***0%***	10%	***10%***	**12.5%**	*17.5%*
1998	**60%**	*42.5%*	*17.5%*	12.5%	7.5%	5%
1999	2.5%	0%	5%	2.5%	5%	5%
2000	**20%**	17.5%	*17.5%*	*20%*	*12.5%*	*12.5%*
2001	0%	2.5%	2.5%	12.5%	7.5%	7.5%
2002	0%	0%	7.5%	***0%***	2.5%	5%
2003	**65%**	**45%**	*17.5%*	*17.5%*	*20%*	**15%**
2004	0%	0%	***7.5%***	7.5%	7.5%	2.5%
2005	0%	*20%*	**15%**	7.5%	*15%*	7.5%
2006	0%	0%	0%	5%	*12.5%*	10%
2007	0%	0%	0%	5%	***5%***	7.5%
2008	2.5%	0%	12.5%	*17.5%*	7.5%	7.5%
2009	5%	*20%*	7.5%	7.5%	*15%*	0%
2010	0%	0%	2.5%	10%	*15%*	*20%*
2011	0%	7.5%	12.5%	**15%**	7.5%	***5%***
2012	*22.5%*	12.5%	12.5%	10%	10%	*17.5%*
2013	0%	0%	7.5%	7.5%	5%	10%
2014	0%	5%	10%	5%	7.5%	*15%*

When the forecast predicts a cold spell, that is, less than a quarter of the total number of predictions (HE + FA = 77 times out of 342 forecasts), the event takes place about 36% of the times. Conversely, when the system foresees no CSPI alert, that is, 265 times (more than 77% of the total), it is almost always right: CSPI occurs once out of 10 predicted nonevents. The outcome, then, differs considerably depending on the forecast, meaning the system contains strong resolution.

Discrimination (and resolution) shown by the MSys indicates that the forecast contains potentially useful information (Mason 2004) that may be exploited by end users and translated into economic value (see section 3c).

Figure 8 shows the geographical distribution of the four possible outcomes in the three aggregated start dates for NCT. The area where CSPI is calculated differs at each start date, forecast week, and ensemble member; the number and location of the enclosed grid points depend on the correlation with the nut farming domain [see Eq. (2)]. Therefore, grid points close to the nut farming sites will be more frequently shown than those located farther away.

To draw the map, we proceeded as follows. For every lead week in which a CSPI is forecasted, and CSPI occurs in ERAI, each grid point of all the MSys ensemble members is flagged as a hit event outcome. For instance, in week 1, HEs take place in 1996 and 2012 on the 1 March start date (Table S4), in 1998, 2000, and 2003 on the 15 March start date (Table 3), and again in 1997, 1998, and 2005 on the 1 April start date (Table S5). Therefore, in week 1, the maximum frequency of HEs for each grid point is 280 (40 ensemble members multiplied by 7 hit events).

We apply the same methodology for every week and each of the four elements of the contingency table, obtaining the outcome frequency (OF) for each element. To illustrate this, we compute the indicator OF in the following way:

$$\mathrm{OF}_{\mathrm{gp},w,e} = \frac{nE_{w,e} \times M_{\mathrm{gp},w}}{M_T \times \mathrm{nSD}}, \tag{15}$$

where $\mathrm{OF}_{\mathrm{gp},w,e}$ is the outcome frequency per grid point (gp) and week w for each element e), $nE_{w,e}$ is the number of outcomes per week for each element, $M_{\mathrm{gp},w}$ is the number of members per grid point and week, M_T is the total number of ensemble members (40), and nSD is

TABLE 4. Contingency tables for the aggregated start dates. Font conventions are as in Table 3.

	Northern coast of Turkey (all start dates)		
		ERAI yes	ERAI no
Weeks 1–2	Model yes	**13**	*7*
	Model no	***7***	87
Weeks 3–4	Model yes	**7**	*20*
	Model no	***12***	75
Weeks 5–6	Model yes	**8**	*22*
	Model no	***11***	73

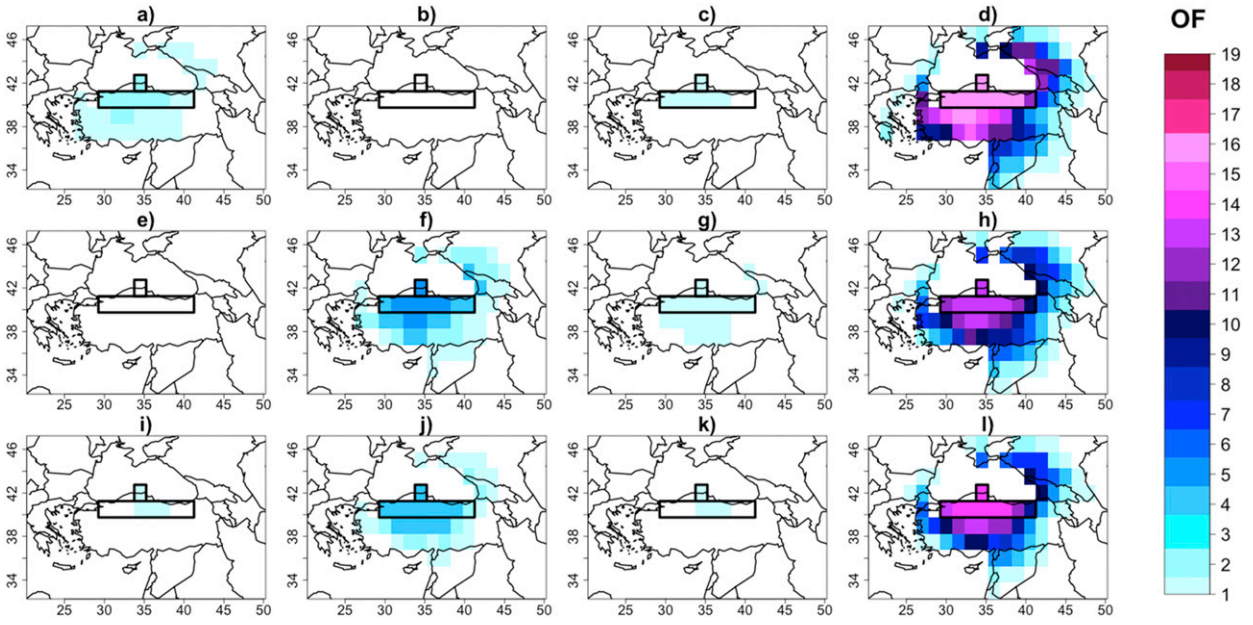

FIG. 8. Geographical distribution of the NCT outcome frequencies, that is, the average occurrence of each of the four elements of the contingency tables [see Eq. (15)]. (a),(e),(i) Hit events (HE), (b),(f),(j) false alarms (FA), (c),(g),(k) missed events (ME), and (d),(h),(l) correct rejections (CR). (top) Forecast weeks 1–2, (middle) forecast weeks 3–4, and (bottom) forecast weeks 5–6.

the number of start dates (three). A grid point showing large values of OF proves the strong correlation between the 2-m temperature variability of that grid point and that of the NCT domain.

As we did in Table 3, results are shown by the two-week aggregation: Figs. 8a–d illustrate the four outcomes (HE, FA, ME, and CR, respectively) for weeks 1–2, Figs. 8e–h illustrate the outcomes for weeks 3–4, and Figs. 8i–l illustrate the outcomes for weeks 5–6.

In the first two weeks, most of the recorded cold spells have been correctly forecasted, as so the noncold spells, hence the HE and CR outcomes have the highest frequency. The reason why frequency of correct rejections is so much higher than that of hit events is because CSPI events are rare.

In weeks 3–4, frequency of CR decreases, because the forecast sometimes predicts a CSPI and the event does not take place; in fact, false alarm frequency increases in turn. At this lead time there is also an increase of missed event frequency, because the forecast misses a few CSPIs, therefore HE frequency declines. A very similar behavior can be seen in weeks 5–6, where the performance is almost identical to that in weeks 3–4.

c. *Value of the forecast*

The potential value of the CSPI forecast for a possible decision-maker is estimated through a simple cost–loss model (Richardson 2000). In practical terms, an agribusiness player may start trading on nut price in advance, when the kernel is not even formed yet, on the basis of the forecast outcome. In the event of a predicted cold spell, they can fix the price beforehand, guaranteeing a net gain in case the cold spell occurs.

Results are shown in Fig. 9 and display the fraction of economic gain potentially imparted by the use of subseasonal forecasts. Again, the contingency tables for the three start dates are merged together. Since we have no information on either the cost C of action (e.g., putting in place a forecast system and acting depending on the predictions) or the potential loss L (i.e., the increase in price following a cold spell occurrence), results are expressed in terms of the C/L ratio. In general, when $C/L \ll 0.1$, the mitigation strategy is so cheap that the decision-maker would always put it into action; therefore, the forecast is not needed or useful. In contrast, when $C/L \lambda 1$, C and L are comparable, so it is not worthwhile to act (i.e., to use the forecast); rather, it is more remunerative to pay for the possible loss.

Since skill is higher at lower lead times, the forecast value of weeks 1–2 is higher than that of weeks 3–6 for both targets. Agribusiness operators in northern coastal Turkey benefit from the use of the medium-range forecast (target up to two weeks), with a potential gain well exceeding 50% for a C/L ratio around 0.2. However, the forecast value is not marginal after week 3, with some potential users obtaining almost 20% gain by using the long-range predictions. Our results indicate that the potential forecast value is similar in the two areas, and

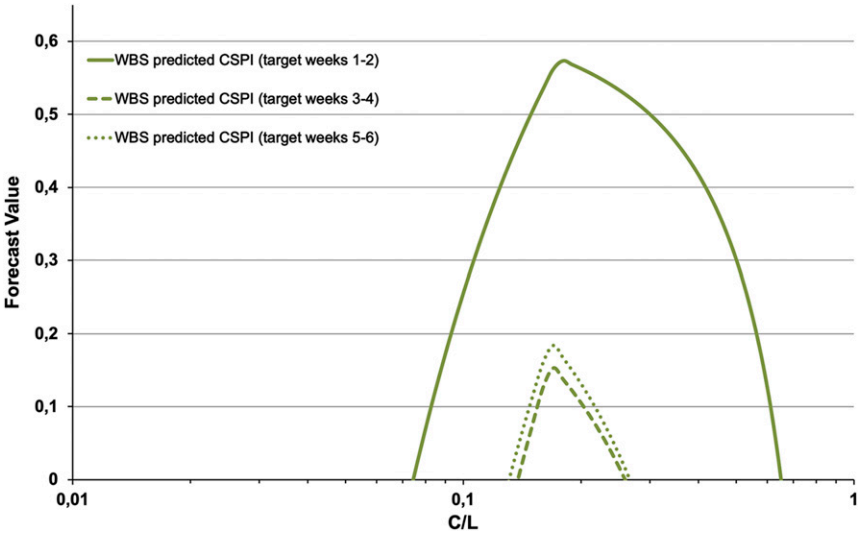

FIG. 9. Value of the forecast (fraction of economic gain imparted by the use of subseasonal forecasts over no use of any forecast, multiplied by 100 to get a percentage) obtained through the cost–loss model. Full lines are for weeks 1–2, dashed lines are for weeks 3–4, and dotted lines are for weeks 5–6.

that there is no significant benefit loss beyond week 3. In fact, the potential value of forecast weeks 3–4 and weeks 5–6 is very much alike.

4. Discussion

This work has been designed to assess the potential usefulness of subseasonal forecasts, treated as an early-warning tool to plan mitigation strategies against cold spells hitting hazelnut production in the Black Sea area. It is not meant to evaluate concrete frost-induced damages that could impact plant tissues or phenology.

Recent spring frost events were recorded by in situ weather stations in 1993, 1997, 2000, 2003, 2004, and 2014 (Erdogan 2018), with ERAI data matching this list except for the 2014 cold spell, which strongly hit only the easternmost side of the domain and is not disclosed in Table 2. Among these events, only some have been reported to cause extensive damage to hazelnut production. Temperatures in north Turkey dropped far below 0°C between March and April 2004 (Ustaoglu 2012), giving rise to serious crop losses: more than 70% of the harvest, with peaks of almost 90%, was wasted compared to the normal production yielded in 2002 (Erdogan and Aygün 2017). As a result, the country's expected production decreased from 600 000 to 350 000 tons, and prices exploded. Another frost event, similar in magnitude but more circumscribed in space, occurred in 2014, when late winter was much warmer than usual and vegetation function started early all throughout the country. On the night of 30 March, heavy snow followed by low temperatures of −5°C destroyed young leaves, shoots, and pollinated female flower clusters, wiping out the crop above the elevation of 300 m in the eastern Black Sea region. Hazelnuts orchards in the western Black Sea region were also damaged but in a less dramatic way, while orchards along the coasts were not affected by the cold spell. Conversely, the other aforementioned frosts were not reported to end up damaging the following hazelnut harvests, despite taking place in the same time of year (Erdogan and Aygün 2017). This fact is surely linked to the duration or the intensity of the cold spells, but also to the physiological state of the plants at the arrival of the frost.

For this sort of consideration, a crop model forced by climate forecasts should be run. In this case, variable thresholds would be chosen for CSPI definition, since a constant 10th percentile limit does not guarantee the onset of a frost, and even less whether this frost is detrimental for the plant. Moreover, higher-resolution data would be needed for this aim, since the S2S grid covers, in one single point, the Black Sea coastal areas and the Pontic Mountain peaks, whose climates are totally incomparable. Finally, model output bias correction should be implemented to take into consideration the MSys systematic errors, and similarly, a cost–loss analysis could be implemented using atmospheric circulation variables [e.g., weather type frequencies of occurrence, which have been shown to be good predictors for extreme rainfall events (Muñoz et al. 2016; Doss-Gollin et al. 2018)] rather than near-surface temperatures directly.

This study does not include any specific analysis to relate the forecast skill to a specific source. Hence,

hypothesizing over possible drivers of predictability would be just speculation. However, the fact that weeks 3–4 and 5–6 have a similar level of skill suggests that predictability could come from lower-frequency variability (possibly antecedent land surface state) rather than higher-frequency signal like the MJO. This conjecture is corroborated by a few studies that linked late winter snow with temperature cold anomalies in eastern Europe (Shongwe et al. 2007), which is shown to be a region of strong snow–atmosphere coupling (Xu and Dirmeyer 2011). Snow, in fact, acts on the atmosphere by both changing the radiatively driven albedo and impacting the hydrological cycles, since soils covered in snow are slowly provided with water able to infiltrate in depth, thus affecting temperatures in the months to come (Xu and Dirmeyer 2013). Such mechanisms are able to impart predictability to the system for a few weeks, although they likely explain a small portion of the total temperature variability in the area. On the other hand, even if recent studies have made first attempts to directly link temperature extremes in eastern Europe with MJO (Seo et al. 2016), a clear teleconnection has not been yet established for Turkey in the beginning of spring.

To release information of use for stakeholders, a probabilistic warning needs to be transformed into a definite decision. To help in making this decision, GSS came to use to set the lowest probability limit, that is, the minimum likelihood required for a cold spell forecast. GSS was not used to assess the benefit imparted by the forecast. Rather, it was required to assign each outcome a score, in order to obtain an LPL. Given the way GSS is constructed, it places an equal penalty on the two forecast ''mistakes'': both false alarms and missed events are, in fact, marked with a −1. Practically, these individual outcomes can have very different consequences from a decision-maker's perspective: a false alarm generates a useless expenditure, but a missed event could lead to a catastrophic loss. In real-world use, such scores should be weighted according to a wider assessment that takes risk analysis into top-level consideration.

5. Conclusions

In this work, a subseasonal forecast multisystem was established, blending four of the forecast systems involved in the Subseasonal to Seasonal Prediction project (Vitart et al. 2017); it was analyzed for three different initializations (1 March, 15 March, and 1 April) over 19 recent springs (1996–2014). The main aims of this study were the evaluation of the multisystem near-surface temperature global forecast quality for boreal spring, the assessment of cold spell prediction in the southern region facing the Black Sea, key for the production of hazelnuts, and the estimation of potential forecast value for end users.

The 2-m temperature forecast was first evaluated using a deterministic metric, the ensemble mean correlation, globally showing that the correlation between the multisystem and reanalysis is higher than that of any single model, especially for longer lead weeks. The equatorial regions, as well as midlatitude highly populated areas such as the Great Lakes region and northeastern China maintain substantial skill at week 5, higher than 0.5, and further studies are planned to explore potential sources of such high skill. Probabilistic metrics such as the ignorance skill score and the generalized ROC score show that although at global scale skill tends to decrease with lead time, some regions seem to exhibit an increase in skill, which is related to overconfidence in the forecasts at shorter lead times, as confirmed by a decomposition of the ignorance skill score. Overall, the tropics show better-than-climatological skill values in near-surface temperatures, even at lead times as large as week 6.

In the south coastal Black Sea, correlations are less remarkable after the deterministic skill time, and generally show values below 0.4 beyond week 2. There is high variability across the three spring start dates as well as across the four models. However, the choice of concatenating the start dates effectively triples the number of forecast years, removing part of the noise, and clearly shows that the multisystem is always the best or the second best choice.

Although the area does not exhibit temperature prediction skill after week 2, this work shows that low skill for 2-m temperatures does not prevent the forecast from being potentially valuable to decision-makers if a different but related variable is used instead. Indeed, previous studies have shown that considering number of rainy or dry days tends to provide higher skill than accumulated rainfall (Moron et al. 2010; Muñoz et al. 2015, 2016), suggesting that frequency-based variables like cold spells are less noisy and thus more predictable. In fact, the CSPI subseasonal forecast is shown to embody resolution and discrimination, which are crucial attributes to determine the information usefulness.

When a cold spell is detected in the reanalysis, the multisystem is able to predict it around half of the time (i.e., 48%). When CSPI does not occur, the system incorrectly foresees it about 17% of the time. The forecast probability is considerably higher when CSPI occurs than when CSPI is not observed, meaning the forecast is different depending on the outcome; hence, the forecast holds discrimination.

In turn, when the multisystem predicts a cold spell, the event takes place about 36% of the time. Conversely, when the system foresees a no-CSPI alert, it is almost

always right: CSPI occurs in only 10% of the predicted nonevents. The outcome, then, differs considerably depending on the forecast, meaning the system contains strong resolution. These two characteristics reveal that the forecast incorporates potentially useful information that should not be ignored by decision-makers.

To the best of our knowledge, a cost–loss model was used for the first time in this context to explore the value of subseasonal predictions applied to cold spells. We found that the potential value of the forecasts is conspicuous for a number of users, who may potentially benefit from the use of subseasonal predictions compared to a no-action strategy. On the northern coast of Turkey, as expected, the confident lead time (weeks 1–2) has more intrinsic value than the dispersive lead times (weeks 3–4 and 5–6). However, even forecasts supplied 3–6 weeks in advance may result in up to a 20% economic gain for agribusiness operators, despite the evident loss in 2-m temperature skill after the deterministic skill time.

Acknowledgments. The multisystem weekly subseasonal forecasts for the three start dates, and the weekly corresponding ERA-Interim values, used to calculate cold spells according to Eq. (1), are available upon request to the main author. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant 730482 (CLARA), and it has been in part conducted within the Project MEDSCOPE, which is part of ERA4CS, an ERA-NET initiative by JPI Climate, and funded by AEMET (ES), ANR (FR), BSC (ES), CMCC (IT), CNR (IT), IMR (BE), and Météo France (FR), with cofunding by the European Union (Grant 690462). AGM was partially supported by NOAA's Modeling, Analysis, Predictions and Projections (MAPP) Award NA18OAR4310275. The research was carried out with the cooperation and contribution of the Hazelnut company division of Ferrero Group.
SM is thankful to Diana C. Rambaldi, Laura Giustarini, Simone Bregaglio, Tommaso De Gregorio, and Stefano Tibaldi for providing insightful comments.
The constructive comments by three anonymous reviewers contributed to substantial improvement of the first version of the manuscript.

APPENDIX

Abbreviations

CR	Correct rejection
CSPI	Cold spell power index
ERAI	ERA-Interim
EXT	Extent of the detected cold spell
FA	False alarm
GSS	Gerrity skill score
HE	Hit event
IGN	Ignorance metric
ISS	Ignorance skill score
LPL	Lowest probability limit
ME	Missed event
MGN	Magnitude of the detected cold spell
MSys	The multisystem used in this study
NCT	Northern coast of Turkey
REL	Reliability of the forecast
RES	Resolution of the forecast
TAC	Temperature anomaly correlation
UNC	Uncertainty of the forecast

REFERENCES

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer, 2011: Evaluation of probabilistic quality and value of the ensembles multimodel seasonal forecasts: Comparison with DEMETER. *Mon. Wea. Rev.*, **139**, 581–607, https://doi.org/10.1175/2010MWR3417.1.

Alexander, L., and Coauthors, 2006: Global observed changes in daily climate extremes of temperature and precipitation. *J. Geophys. Res.*, **111**, D05109, https://doi.org/10.1029/2005JD006290.

Ardilouze, C., L. Batté, and M. Déqué, 2017: Subseasonal-to-seasonal (S2S) forecasts with CNRM-CM: A case study on the July 2015 west-European heat wave. *Adv. Sci. Res.*, **14**, 115–121, https://doi.org/10.5194/asr-14-115-2017.

Batté, L., C. Ardilouze, and M. Déqué, 2018: Forecasting West African heat waves at subseasonal and seasonal time scales. *Mon. Wea. Rev.*, **146**, 889–907, https://doi.org/10.1175/MWR-D-17-0211.1.

Beyhan, N., and F. Odabas, 1996: The investigation of compatibility relationships of some hazelnut cultivars. *Acta Hortic.*, **445**, 173–178, https://doi.org/10.17660/ActaHortic.1997.445.23.

Buizza, R., 2008: Comparison of a 51-member low-resolution (TL399L62) ensemble with a 6-member high-resolution (TL799L91) lagged-forecast ensemble. *Mon. Wea. Rev.*, **136**, 3343–3362, https://doi.org/10.1175/2008MWR2430.1.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

DeFlorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. M. Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *J. Hydrometeor.*, **19**, 409–426, https://doi.org/10.1175/JHM-D-17-0135.1.

DelSole, T., and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678, https://doi.org/10.1175/MWR-D-14-00045.1.

Doblas-Reyes, F. J., M. Déqué, and J.-P. Piedelievre, 2000: Multimodel spread and probabilistic seasonal forecasts in provost. *Quart. J. Roy. Meteor. Soc.*, **126**, 2069–2088, https://doi.org/10.1256/smsqj.56704.

Doss-Gollin, J., Á. G. Muñoz, S. J. Mason, and M. Pastén, 2018: Heavy rainfall in Paraguay during the 2015/16 austral summer: Causes and subseasonal-to-seasonal predictive skill. *J. Climate*, **31**, 6669–6685, https://doi.org/10.1175/JCLI-D-17-0805.1.

Erdogan, V., 2018: Hazelnut production in Turkey: Current situation, problems and future prospects. *Acta Hortic.*, **1226**, 13–24, https://doi.org/10.17660/ActaHortic.2018.1226.2.

——, and A. Aygün, 2017: Late spring frosts and its impact on Turkish hazelnut production and trade. Nucis Newsletter, No. 17, FAO CIHEAM, Constantí, Spain, 25–27.

FAOSTAT, 2016: Statistics division. Economic and Social Development Department, accessed 12 March 2019, http://www.fao.org/faostat/en/#data/QC.

Ferrone, A., D. Mastrangelo, and P. Malguzzi, 2017: Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Adv. Sci. Res.*, **14**, 123–129, https://doi.org/10.5194/asr-14-123-2017.

Furtado, J., J. Cohen, and E. Tziperman, 2016: The combined influences of autumnal snow and sea ice on Northern Hemisphere winters. *Geophys. Res. Lett.*, **43**, 3478–3485, https://doi.org/10.1002/2016GL068108.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370, https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Gerrity, J. P., Jr., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709–2712, https://doi.org/10.1175/1520-0493(1992)120<2709:ANOGAM>2.0.CO;2.

Guo, Z., P. A. Dirmeyer, and T. DelSole, 2011: Land surface impacts on subseasonal and seasonal predictability. *Geophys. Res. Lett.*, **38**, L24812, https://doi.org/10.1029/2011GL049945.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2006: DEMETER and the application of seasonal forecasts. *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 674–692.

Hoell, A., M. Barlow, M. C. Wheeler, and C. Funk, 2014: Disruptions of El Niño–southern oscillation teleconnections by the Madden–Julian Oscillation. *Geophys. Res. Lett.*, **41**, 998–1004, https://doi.org/10.1002/2013GL058648.

Koster, R., and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.*, **12**, 805–822, https://doi.org/10.1175/2011JHM1365.1.

Krishnamurti, T. N., A. Chakraborty, R. Krishnamurti, W. K. Dewar, and C. A. Clayson, 2006: Seasonal prediction of sea surface temperature anomalies using a suite of 13 coupled atmosphere–ocean models. *J. Climate*, **19**, 6069–6088, https://doi.org/10.1175/JCLI3938.1.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86, https://doi.org/10.1214/aoms/1177729694.

Li, S., and A. W. Robertson, 2015: Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, **143**, 2871–2889, https://doi.org/10.1175/MWR-D-14-00277.1.

Liang, P., and H. Lin, 2018: Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. *Climate Dyn.*, **50**, 1007–1022, https://doi.org/10.1007/s00382-017-3658-1.

Lorenz, E. N., 1963: Section of planetary sciences: The predictability of hydrodynamic flow. *Trans. N. Y. Acad. Sci.*, **25**, 409–432, https://doi.org/10.1111/j.2164-0947.1963.tb01464.x.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Mason, S. J., 2004: On using "climatology" as a reference strategy in the brier and ranked probability skill scores. *Mon. Wea. Rev.*, **132**, 1891–1895, https://doi.org/10.1175/1520-0493(2004)132<1891:OUCAAR>2.0.CO;2.

——, and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **137**, 331–349, https://doi.org/10.1175/2008MWR2553.1.

——, and M. K. Tippett, 2019: Climate Predictability Tool version 15.7.11. Columbia University, https://doi.org/10.7916/d8-kb0s-2816.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

Moron, V., A. W. Robertson, and M. N. Ward, 2010: Seasonal predictability and spatial coherence of rainfall characteristics in the tropical setting of Senegal. *Mon. Wea. Rev.*, **134**, 3248–3262, https://doi.org/10.1175/MWR3252.1.

Muñoz, Á. G., L. Goddard, A. W. Robertson, Y. Kushnir, and W. Baethgen, 2015: Cross–time scale interactions and rainfall extreme events in southeastern South America for the austral summer. Part I: Potential predictors. *J. Climate*, **28**, 7894–7913, https://doi.org/10.1175/JCLI-D-14-00693.1.

——, ——, S. J. Mason, and A. W. Robertson, 2016: Cross–time scale interactions and rainfall extreme events in southeastern South America for the austral summer. Part II: Predictive skill. *J. Climate*, **29**, 5915–5934, https://doi.org/10.1175/JCLI-D-15-0699.1.

——, X. Yang, G. A. Vecchi, A. W. Robertson, W. F. Cooke, 2017: A weather-type-based cross-time-scale diagnostic framework for coupled circulation models. *J. Climate*, **30**, 8951–8972, https://doi.org/10.1175/JCLI-D-17-0115.1.

——, C. Coelho, A. W. Robertson, and S. Mason, 2018: How much can model output statistics improve sub-seasonal predictive skill? *Second Int. Conf. on Subseasonal-to-Seasonal Prediction*, Boulder, CO, WCRP, A3-03, https://www.wcrp-climate.org/images/WCRP_conferences/S2S_S2D_2018/pdf/Programme/orals/presentations/A3-03_AngelGMunoz_MAC.pdf.

Orsolini, Y., R. Senan, G. Balsamo, F. Doblas-Reyes, F. Vitart, A. Weisheimer, A. Carrasco, and R. Benestad, 2013: Impact of snow initialization on sub-seasonal forecasts. *Climate Dyn.*, **41**, 1969–1982, https://doi.org/10.1007/s00382-013-1782-0.

Palmer, T., Č. Branković, and D. Richardson, 2000: A probability and decision-model analysis of provost seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.*, **126**, 2013–2033, https://doi.org/10.1256/smsqj.56702.

Peings, Y., J. Cattiaux, and H. Douville, 2013: Evaluation and response of winter cold spells over western Europe in CMIP5 models. *Climate Dyn.*, **41**, 3025–3037, https://doi.org/10.1007/s00382-012-1565-z.

Piedelievre, J., 2000: Numerical seasonal predictions using global climate models. researches on the long range forecasting at Météo France. *Stochastic Environ. Res. Risk Assess.*, **14**, 319–338, https://doi.org/10.1007/PL00013451.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Rodrigo, J., 2000: Spring frosts in deciduous fruit trees–Morphological damage and flower hardiness. *Sci. Hortic.*, **85**, 155–173, https://doi.org/10.1016/S0304-4238(99)00150-8.

Rodwell, M. J., and F. J. Doblas-Reyes, 2006: Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Climate*, **19**, 6025–6046, https://doi.org/10.1175/JCLI3944.1.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660, https://doi.org/10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Seo, K.-H., H.-J. Lee, and D. M. Frierson, 2016: Unraveling the teleconnection mechanisms that induce wintertime temperature anomalies over the Northern Hemisphere continents in response to the MJO. *J. Atmos. Sci.*, **73**, 3557–3571, https://doi.org/10.1175/JAS-D-16-0036.1.

Shongwe, M. E., C. A. T. Ferro, C. A. S. Coelho, and G. Jan van Oldenborgh, 2007: Predictability of cold spring seasons in Europe. *Mon. Wea. Rev.*, **135**, 4185–4201, https://doi.org/10.1175/2007MWR2094.1.

Siebert, A., 2016: Analysis of index insurance potential for adaptation to hydroclimatic risks in the West African Sahel. *Wea. Climate Soc.*, **8**, 265–283, https://doi.org/10.1175/WCAS-D-15-0040.1.

Thomas, J. A., A. A. Berg, and W. J. Merryfield, 2016: Influence of snow and soil moisture initialization on sub-seasonal predictability and forecast skill in boreal spring. *Climate Dyn.*, **47**, 49–65, https://doi.org/10.1007/s00382-015-2821-9.

Tripathi, O. P., A. Charlton-Perez, M. Sigmond, and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environ. Res. Lett.*, **10**, 104007, https://doi.org/10.1088/1748-9326/10/10/104007.

Ustaoglu, B., 2012: The effect of climatic conditions on hazelnut (Corylus avellana) yield in Giresun (in Turkish). *Marmara Coğrafya Derg.*, **26**, 302–323.

Vigaud, N., A. W. Robertson, M. K. Tippett, and N. Acharya, 2017: Subseasonal predictability of boreal summer monsoon rainfall from ensemble forecasts. *Front. Environ. Sci.*, **5**, 67, https://doi.org/10.3389/fenvs.2017.00067.

Vitart, F., and F. Molteni, 2010: Simulation of the Madden–Julian Oscillation and its teleconnections in the ECMWF forecast system. *Quart. J. Roy. Meteor. Soc.*, **136**, 842–855, https://doi.org/10.1002/qj.623.

——, 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.

Weijs, S. V., R. Van Nooijen, and N. Van De Giesen, 2010: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Mon. Wea. Rev.*, **138**, 3387–3399, https://doi.org/10.1175/2010MWR3229.1.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Xu, L., and P. Dirmeyer, 2011: Snow-atmosphere coupling strength in a global atmospheric model. *Geophys. Res. Lett.*, **38**, L13401, https://doi.org/10.1029/2011GL048049.

——, and ——, 2013: Snow–atmosphere coupling strength. Part II: Albedo effect versus hydrological effect. *J. Hydrometeor.*, **14**, 404–418, https://doi.org/10.1175/JHM-D-11-0103.1.