Skill and Reliability of Seasonal Forecasts for the Chinese Energy Sector

PHILIP E. BETT, HAZEL E. THORNTON, AND JULIA F. LOCKWOOD

Met Office Hadley Centre, Exeter, United Kingdom

ADAM A. SCAIFE

Met Office Hadley Centre, and College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

NICOLA GOLDFING AND CHRIS HEWITT

Met Office Hadley Centre, Exeter, United Kingdom

RONG ZHU AND PEIQUN ZHANG

Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing, China

CHAOFAN LI

Center for Monsoon System Research, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

(Manuscript received 17 March 2017, in final form 15 August 2017)

ABSTRACT

The skill and reliability of forecasts of winter and summer temperature, wind speed, and irradiance over China are assessed using the Met Office Global Seasonal Forecast System, version 5 (GloSea5). Skill in such forecasts is important for the future development of seasonal climate services for the energy sector, allowing better estimates of forthcoming demand and renewable electricity supply. It was found that, although overall the skill from the direct model output is patchy, some high-skill regions of interest to the energy sector can be identified. In particular, winter mean wind speed is skillfully forecast around the coast of the South China Sea, related to skillful forecasts of the El Niño–Southern Oscillation. Such information could improve seasonal estimates of offshore wind-power generation. In a similar way, forecasts of winter irradiance have good skill in eastern central China, with possible use for solar-power estimation. Skill in predicting summer temperatures, which derives from an upward trend, is shown over much of China. The region around Beijing, however, retains this skill even when detrended. This temperature skill could be helpful in managing summer energy demand. While both the strengths and limitations of the results presented here will need to be considered when developing seasonal climate services in the future, the outlook for such service development in China is promising.

1. Introduction

The energy sector has long been a key user of weather and climate information across time scales: short-range weather forecasts (e.g., Taylor and Buizza 2003; Costa et al. 2008), longer-range forecasts out to several weeks ahead (e.g., Dubus 2014), and projections of possible future climates decades ahead (e.g., McColl et al. 2012; Wang et al. 2014). These forecasts are all used to inform the planning, development, management, and running of energy systems on those time scales. The energy sector has also been a leader in demonstrating demand for seasonal-to-decadal climate prediction (Buontempo et al. 2010; Bruno Soares and Dessai 2015). In particular, seasonal forecasts of the climate in the coming 3-month period have the potential for providing real added value, in both the practical sense and the financial sense, across a range of areas within the energy sector (Troccoli 2010; Doblas-Reyes et al. 2013; Dessai and

Corresponding author: Philip E. Bett, philip.bett@metoffice.gov.uk

DOI: 10.1175/JAMC-D-17-0070.1

© 2017 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
Energy demand is strongly related to air temperature (e.g., Valor et al. 2001; Hor et al. 2005; Apadula et al. 2012; Zhang et al. 2014; Thornton et al. 2016), and the potential use of seasonal climate forecasting in demand management has been recognized for many decades (e.g., Brown Weiss 1982 and references therein). The need to reduce both greenhouse gas emissions and air pollution has driven an increase in the amount of electricity supplied by renewable sources. This has, in turn, resulted in an increase in the weather dependence of energy supply systems and therefore an increase in the possible utility of weather and climate forecasting to the sector.

The energy sector in China faces more severe issues than do the energy sectors of other countries. China is the coldest country during winter, and the hottest during summer, among countries located at the same latitudes (excluding desert regions), and therefore the demands for heating in winter and cooling in summer are high. Under the background of global climate change, extreme climate events have changed significantly in China over the past 60 years. The increase in the number of high-temperature and drought days (Qin et al. 2015) has exacerbated the energy gap for cooling and irrigation when such extremes occur. Other particular features are 1) the increases in demand that are due to rapid urbanization (Wang 2014; Lin and Ouyang 2014) and 2) the recent large growth in both installed and planned renewable energy capacity (e.g., Hong et al. 2013; CNREC 2014; Lo 2014; Qiang et al. 2016). Any shift in the future from coal-driven combined heat and power facilities to more electricity-based building heating would also have significant impacts on the use of renewable energy sources (e.g., Zhang et al. 2016).

If skillful, seasonal forecasts would be very useful to the Chinese energy sector. For example, seasonal forecasts could give an early warning of a season of high demand such as a particularly cold winter or hot summer, or of reduced supply as a result of low wind speeds, or more cloudy/hazy periods that reduce solar-power generation. Predicting conditions that could damage energy infrastructure, such as storms, could also be valuable. In all of these cases, seasonal forecasts could allow improved mitigation plans to be put in place: rescheduling maintenance work, making early decisions around staff availability, and financial planning for the coming 3-month period. The information could be used by a range of people, including industry regulators, network operators, energy-production companies, maintenance contractors, and traders in the financial markets. In many cases, using a seasonal forecast might require changes in an organization’s decision-making process, however.

Seasonal forecasts are most useful if they have sufficient skill to allow decision-making. Furthermore, what “sufficient” means (beyond being statistically significant) will depend on the particular use case. If forecasts are not sufficiently skillful, then, although they might be happy to receive forecast information, organizations might not be able to make a decision that is based on that information.

Although skillful forecasts for some variables in some parts of the world have been possible for some time (e.g., Arribas et al. 2011), recent advances in seasonal forecasting systems have led to major improvements in the skill of extratropical features such as the North Atlantic Oscillation (NAO; e.g., Athanasiadis et al. 2014, 2016; Butler et al. 2016; Smith et al. 2016, and references therein). Scaife et al. (2014) demonstrated skill in NAO forecasts from version 5 of the Met Office’s Global Seasonal Forecasting System (GloSea5; MacLachlan et al. 2015). This has led to the development of seasonal climate services for various sectors in the United Kingdom, including hydrology (Svensson et al. 2015), transport (Palin et al. 2016), and energy (Clark et al. 2017). Seasonal forecasting has a long history in China, but the traditional low skill from dynamical models has led to a wide quantity of literature in statistical downscaling and statistical forecasting (e.g., Xiao et al. 2012; Wang et al. 2015; Xing et al. 2016; Golding et al. 2017, and references therein). Much of this literature has focused on forecasting the monsoon and associated precipitation (e.g., Ke et al. 2011; Ying and Ke 2012; Tung et al. 2013; Peng et al. 2014; Wu and Yu 2016). Development of wind speed predictions has focused on short-term (weather) forecasts, although there has been some work done on the longer-term monthly time scale (Guo et al. 2011). It is timely therefore to examine the skill of GloSea5 in China for direct seasonal forecasts of energy-relevant climate variables. The results could allow the development of future climate services (Golding et al. 2017), that is, the provision and use of climate information to enable better-informed decisions.

The development and use of such climate services has become a major undertaking worldwide, with international coordination being facilitated by the Global Framework for Climate Services (GFCS; Hewitt et al. 2012). The GFCS focuses on five priority sectors, one of which is energy.
which is the energy sector. China is developing its own framework aligned to the GFCS, called the China Framework for Climate Services (http://www.cma.gov.cn/en2014/20150311/2015323e/2015323e07/2015323e09/201503/20150317_276866.html; Golding et al. 2017). It brings together actors involved in scientific research and climate-service development with service providers and users to ensure that available capability and services meet users’ needs.

In this paper we assess the skill of seasonal forecasts of wind speed, irradiance, and temperature across China from GloSea5 and consider the implications for the wind-power, solar-power, and energy-demand sectors. We first describe the datasets and methods used in section 2. We then present our results in section 3, considering a China-wide overview of each variable before focusing on some specific areas of interest. We discuss our conclusions in section 4.

2. Data and analysis methods

a. Datasets

In this paper we use the hindcast dataset produced to assess the version of GloSea5 that was deployed operationally at the Met Office in February of 2015. It is based on the second Global Coupled configuration (GC2) of the third Hadley Centre Global Environmental Model (HadGEM3), described in Williams et al. (2015). HadGEM3-GC2 uses the “GA6.0” Global Atmosphere configuration of the Met Office Unified Model (version 8.4) as its atmospheric component, on an “N216” grid (a horizontal resolution of 0.83° in longitude and 0.55° in latitude) and 85 vertical levels reaching a height of 85 km near the mesopause (Walters et al. 2017). This is coupled to the “GL6.0” Global Land configuration of the Joint U.K. Land Environment Simulator (JULES) land surface model (Best et al. 2011), the “GO5.0” Global Ocean configuration of the Nucleus for European Modelling of the Ocean (NEMO) model with a 0.25° nominal resolution and 75 vertical levels (version 3.4; Megann et al. 2014; Madec 2008), and the “GSf6.0” Global Sea Ice configuration of the Los Alamos Sea Ice Model (CICE, version 4.1; Rae et al. 2015; Hunke and Lipscomb 2010). GloSea5 is described in full in MacLachlan et al. (2015) and references therein.

The assessment hindcast was produced to examine the skill of the system in forecasting for winter (December–February: DJF) and summer (June–August: JJA) only.

Lagged ensemble “forecasts” were produced by collating eight-member ensemble forecasts initialized on 25 October, 1 November, and 9 November for DJF and on 25 April, 1 May, and 9 May for JJA. This method yields a total of 24 members for each hindcast season. The DJF hindcasts cover from (boreal) winter 1992/93 to winter 2011/12, and the JJA hindcasts cover (boreal) summers 1992–2011. The details of the initialization are described in MacLachlan et al. (2015).

Our ability to robustly assess forecast skill is limited by the size of the hindcast, both in terms of the number of years and the number of members. The operational configuration (the forecast version) of GloSea5 uses 42 members, rather than the 24 available in the hindcast used here, and therefore the probability distributions inferred by using the hindcast will be less well resolved than they would be operationally. Furthermore, it has been shown that the skill itself depends directly on the size of the ensemble, because it allows better identification of predictable signals (Scaife et al. 2014; Eade et al. 2014; Dunstone et al. 2016; Li et al. 2016; Lu et al. 2017), although the size of this effect in GloSea5 is less strong in China than in the extratropical North Atlantic Ocean. For our purposes, the ensemble size means that we can regard the levels of skill that are shown here, where significant, to be lower limits on the actual skill that could be realized in the operational system. The robustness of the skill estimates is also limited by the number of years in the hindcast (Kumar 2009), because this places a restriction on the number of different types of event that are sampled in the period of study. The impact of the limited hindcast period is quantified by assessing the statistical significance of the correlations between the hindcast and observations, as described in the next section.

We use the ERA-Interim reanalysis data (ERAI; Dee et al. 2011) as a proxy for observations in this paper. ERAI and other reanalyses are frequently used to validate temperature and wind data from climate models, but their use as a proxy for irradiance observations is more contested (e.g., Boilley and Wald 2015). We compared some of our irradiance results with the satellite-derived observational Surface Solar Radiation Data Set—Heliosat based on Meteosat-East (SARAH-E; Huld et al. 2016; Amillo et al. 2014) and found that, for the seasonal means averaged over the large areas that we consider here and in the standardized units that we use, ERAI compares very well. The use of climatological aerosols in both ERAI and SARAH-E, and indeed in GloSea5, means that the impact of aerosols on interannual variability remains an important uncertainty.

1 That is, 432 cells east–west × 324 cells north–south.
b. Skill-assessment method

In this paper we focus on three meteorological variables of interest to the energy sector in China: near-surface air temperature (related to energy demand), 10-m wind speed (linked to wind-power generation), and downwelling shortwave irradiance at the surface (related to solar-power generation). Precipitation is also of great importance for the energy sector, because China has a very large, and growing, hydroelectric industry. Li et al. (2016) have already shown that GloSea5 has significant skill in forecasting summer precipitation in the Yangtze River basin, where the Three Gorges Dam is located, and this has led to the development of a trial forecast service (Golding et al. 2017). Further work by Lu et al. (2017) found high levels of skill in GloSea5 forecasts of winter precipitation over southeastern China.

We take the following approach to skill assessment. For each variable, we first map the Pearson correlation between the hindcast ensemble mean and the observations. Although the limited time span and ensemble size of the hindcast means that forecasts from single grid cells (or even small regions) are likely to be very noisy and not robust, these maps give a good indicative overview and provide context when selecting larger geographical areas of interest for subsequent analysis. These regions are selected on the basis of their interest to the energy sector: we consider the current and likely future development of substantial energy supply or demand and whether they appear to have some promising skill (regions are not selected on the basis of skill maps alone).

We then assess the skill in each region in more detail. We consider three types of plots, each with an associated skill score:

1) standardized time series (using the hindcast ensemble mean), with the Pearson correlation $r$ [we also estimate the uncertainty in $r$ (95% confidence intervals) using the Fisher $z$ transformation],

2) reliability and sharpness diagrams, with the Brier skill score (BSS), which show the joint distribution of hindcast probabilities and observed frequencies for a particular class of event (the BSS measures how much better the forecast system is relative to use of climatological means in that case) (see appendix A for further details), and

3) “relative operating characteristic” (ROC) diagrams, with the ROC skill score (ROCSS), which describe the ability of the forecast system to distinguish between events occurring or not occurring (see appendix B for further details).

In all cases, we follow the standard WMO (2010) procedure for assessing such forecasts. In particular, this means weighting by the cosines of the grid-cell latitudes when aggregating grid cells in the region in question.

It is useful to have an indication on the correlation maps of where the skill might be significantly nonzero. We use the Fisher $z$ transformation² to draw contours around areas that would be significantly nonzero at the 5% level.³ Note that we are not correcting for multiple testing here, and therefore it should be expected that some of the regions marked as notionally significant on maps will be false positives; the significance contour should not be taken as definitive.

There are some cases in which there is a clear trend running through the data. The reproduction of such a trend by the hindcast is a genuinely useful aspect of skill, because it shows that the forecast system is capable of maintaining the impact of whatever forcing caused the trend, after being initialized. It can, however, hide information about the ability of the model to evolve correctly away from its initialized state more generally, which is also of interest when assessing the model. We have therefore also looked at the correlation skill after detrending, which we perform by simply removing the linear least squares regression fits to the hindcast ensemble mean and observational time series, separately. Note that we are not making any assumptions as to the cause or significance of any trends. The time series is sufficiently short that natural interannual and decadal-scale variability will be very important, even before considering anthropogenic drivers of climate change such as

² Wind-turbine hub heights are around 80–100 m. The skill measures that we use here focus on the relative changes to each variable, however, and on the time scales in which we are interested—seasonal means—there is no significant difference between relative changes in wind speeds at 10 and 100 m.

³ By “standardized” we mean that we subtract the average and then divide by the standard deviation $\sigma$. The result is an anomaly time series in units of $\sigma$. We use the 20-yr period of the hindcast to determine the long-term average and $\sigma$, for both the observations and the hindcast itself.

⁴ For this test to apply exactly, the data must be Gaussian and independent. In our case, the central limit theorem means that this will be a reasonable approximation for our seasonal-average data and that the test will provide a good indication of nonzero correlation.

⁵ In our case of having 20 years of data, the threshold in correlation corresponding to the 5% significance level is $|r| > 0.44$. 
carbon dioxide emissions, land-use change (affecting effective surface roughness and hence wind speed), and aerosol emissions (affecting surface irradiance and temperature). We simply remove the empirical linear trend.

Our reliability and ROC diagrams are made in terms of probabilistic forecasts of particular types of event: we consider the probability of the variable in question being above the median; in the upper, middle, or lower tercile; or in the top or bottom quintile of its historical distribution. These quantiles are calculated for the hindcast and observational datasets independently, from their own climatological characteristics (defined by the period covered by the hindcast). This means that our reliability diagrams and Brier skill scores are insensitive to a simple bias in the mean state between the two datasets.

For each type of event (e.g., upper tercile), the distribution of ensemble members each year provides

---

6 Although ERAI uses climatological aerosols, it assimilates other observations that could be forced by anomalous aerosol emissions, making it particularly complicated to assess in this way.

7 We use cross validation when calculating the quantile values, with a window length of 1 yr (i.e., one DJF or JJA period), following the method of WMO (2010): the quantile in question is calculated separately for each year from the 19 years of data remaining after that year is masked out.
the forecast probability of that event occurring, in each grid cell. Using bins of probability with width 0.1, we then consider all of the years in which the event was forecast to occur with probability in a given bin and count the frequency of times that it was observed to actually occur. The counting is done in each grid cell, and we then pool the counts from all of the grid cells in the region using cosine-of-latitude-weighted sums to calculate the skill scores and reliability/ROC plots.

Although the quantities we consider here are insensitive to a mean bias, there are possible cases in which absolute values can be important. Transfer functions (power curves) for relating meteorological quantities...
to electrical power output are usually threshold based (e.g., Bett and Thornton 2016 and references therein). It remains to be seen if such threshold-based transformations are necessary in a seasonal forecasting context, but in these cases the hindcast climatological values can be used for bias correction (Arribas et al. 2011; MacLachlan et al. 2015).

3. Results

Here, we show maps of the correlation between GloSea5 and ERAI for each variable. In each case, we then go on to examine particular regions in more detail through their regional time series, reliability, and ROC diagrams.

### a. Wind speed

Maps of the correlation between ERAI and the GloSea5 hindcast for 10-m wind speeds are shown in Fig. 1 for winter and summer. Because there are no significant trends in the wind speeds over the hindcast period, maps of the correlation of detrended time series are practically indistinguishable (not shown).

Although there are some areas of significant positive skill for wind speed in DJF, they are patchy. A major highlight, however, is the very high skill in the South China Sea, off the southern and southeastern coasts of China, with some skill being retained inland. This good performance is likely to be related to the skill in forecasting the El Niño–Southern Oscillation ($r = 0.9$ for the...
DJF Niño-3.4 index (see MacLachlan et al. 2015) and its teleconnections over China; these are shown in Fig. 2 in terms of correlations between the Niño-3.4 index and wind speed. The overall response in the region differs in detail between GloSea5 and ERAI, but the significant anticorrelation in the South China Sea is present in both, with weaker wind speeds correlated with El Niño events. In this region, as part of the East Asian winter monsoon, there is strong northeasterly flow around the southeastern edge of the Siberian–Mongolian high pressure feature (Chang et al. 2006). During El Niño events, there is increased subsidence over the Maritime Continent region of Southeast Asia (Ramage 1968), increasing the surface pressure over that region and thus reducing the land–sea pressure gradient and resulting monsoonal winds (Zhang et al. 1996).

Figure 3 shows skill and reliability for winter wind speeds in this South China Sea region (the southeastern green box in Fig. 1). The deterministic ensemble-mean forecast has a correlation of $r \approx 0.8$, and the reliability diagrams show that this region also exhibits skillful, reliable, and sharp probabilistic forecasts of above-median wind speed events. Results are also very good for upper- and lower-tercile events and for the upper and lower quintiles, although these latter are more noisy. This result is reflected in the sharpness diagrams: high-probability forecasts of outer-quintile events in particular are very poorly sampled by the hindcast, whereas forecasts of above-median events are well sampled across the full range of probabilities. It is also important to note that this skill in wind speed is not retained in summer (as seen in the bottom panel of Fig. 1); in that case, the skill scores of $r = 0.13$, BSS = $-0.07$, and ROCSS = 0.004 are not significantly different from zero (not shown).

Some other regions also appear to have reasonably high levels of skill. Figure 4 shows the time series for winter wind speed in north-central China and southern Mongolia (the northern green box in Fig. 1). This is a region of particularly high wind resource (e.g., CNREC 2014; Davidson et al. 2016), and so being able to forecast it could be of great practical use. The skill in this particular region is marginal, however ($r = 0.42$; although the 95% confidence interval just crosses zero, we expect this to be an underestimate of the true forecast skill, as discussed in section 2). Research is ongoing to understand the origin of this skill, but current results suggest that it is related to the model’s skill in predicting the upper-level Middle Eastern jet stream and, to some extent, the Arctic Oscillation. It has previously been demonstrated that these features are connected to winter climate over China (e.g., Zuo et al. 2015; Yang et al. 2004; He et al. 2017), and research is under way to determine the physical mechanism.

Yunnan Province, in southern China, shows modest but significant skill for wind speeds in both winter and summer (Fig. 5; the region is also marked in Fig. 1). Yunnan is very mountainous, and energy production has traditionally been dominated by hydroelectricity. In recent years there has been a substantial drive to utilize the available wind resource (Liang et al. 2015), however, and a seasonal forecast could prove useful.

The wind forecast skill across most of China in summer is indistinguishable from zero. Wind industry developers should be aware of these limitations in skill and
focus on using seasonal forecasts for winter decision-making, in skillful regions.

b. Irradiance

Figure 6 shows correlation maps for irradiance in winter and summer. In winter, there is a broad area of promising skill in eastern China and the East China Sea. This result bears a strong resemblance to the patterns of skill in winter precipitation shown by Lu et al. (2017), perhaps unsurprisingly, because both rainfall and irradiance are strongly related to cloudiness. Lu et al. (2017) determined that the key drivers of precipitation predictability here are 1) ENSO and 2) rainfall in the eastern Indian Ocean/Bay of Bengal. It is reasonable to assume that the same processes that affect winter precipitation in this region also affect cloudiness and therefore downwelling shortwave irradiance at the surface.

FIG. 7. Skill and reliability assessment of winter irradiance for the easternmost coastal region of China marked with a green box in Fig. 6. As in Fig. 3, we (top) show the standardized time series using the ensemble mean and also show (middle) reliability and sharpness diagrams and (bottom) ROC plots for (left) above-median, (center) tercile, and (right) outer-quintile events. Each panel shows its corresponding skill score, and further details on the reliability and ROC plots are given in appendixes A and B, respectively.

Unauthenticated | Downloaded 07/31/22 04:18 PM UTC
Figure 7 shows the winter irradiance skill in eastern China in more detail, using the region in the eastern green box marked in Fig. 6: probabilistic forecasts remain reasonably skillful and reliable for tercile and outer-quintile events. While this is not a region of particularly high solar radiation resource within China (CNREC 2014), it is an area of high population density and many urban centers, including Shanghai. The potential for high levels of demand modulated by large numbers of roof-mounted solar panels means that being able to forecast winters with more or less solar generation than usual could be of value.

In summer, the correlation overall across China is much poorer, with the region that we considered above in eastern China now having $r = 0.03$, which is consistent with zero (not shown). Regions farther west appear to have higher levels of skill, but the skill is still very patchy.

As already discussed, because both ERAI and GloSea5 use climatological aerosols, we are unable to assess the impact of interannual aerosol variability on seasonal irradiance forecasts. This might have a strong impact in urban areas that are affected by haze, for example, and it remains an important uncertainty when considering the application of these results to solar-power generation.

c. Temperature

Figure 8 shows correlation maps for temperature, including a comparison with detrended data. It is clear that there is significant skill in predicting summer temperatures over large areas of China but that in many regions these areas of skill are due in large part to the model reproducing the observed positive trends over the hindcast period; in such cases, using a climate model for seasonal forecasting may not be necessary. In other areas, the skill persists after the trend is removed. The winter temperatures are less affected by trends and, indeed, show very little skill overall. As with winter wind speed, research is under way to improve the forecasts through the use of statistical models that are based on larger-scale atmospheric drivers.

One exception is Yunnan Province in south-central China, which we also highlighted for wind speed skill: here the province shows positive skill for both winter
and summer temperatures, before and after detrending. Yunnan is less urban than many more-eastern parts of China, and therefore the utility of a temperature forecast in energy-demand planning is more limited. Agriculture and tourism are both very important for Yunnan, which could thus still benefit from a skillful seasonal temperature forecast.

It is clear that the potential for useful seasonal forecasts of energy demand and hence temperature is greatest in urban centers. In particular, energy demand in Beijing is strongly related to temperature in summer (Zhang et al. 2014). It is important therefore that the region around Beijing (the northern green box in Fig. 8) also shows some skill for summer temperatures, before and after detrending, with correlations of ~0.5–0.6 (Fig. 9). There is less skill, however, for probabilistic forecasts that are more detailed than the simple “above average” case (note that the trend seen in the time series will also make a positive contribution to the reliability), although this skill might be improved by looking at a larger region and thus reducing statistical noise.

4. Discussion and conclusions
Our results have shown that, although overall skill for energy-relevant variables in China remains patchy,
some specific areas have significant skill: winter wind speeds in the South China Sea (reflecting a known relationship with ENSO), winter solar irradiance in eastern/southern China, and summer temperatures across much of China (due to the trend), including Beijing (even when detrended). Taken together with similarly promising results for skillful summer precipitation forecasts in the Yangtze River basin (Li et al. 2016) and winter precipitation forecasts in southeastern China (Lu et al. 2017), there are clear opportunities to develop useful seasonal climate services for specific cases within China. Indeed, taking our results and those on precipitation skill together, there are clear potential climate-service applications beyond the energy sector: for example, forecasting risks to agriculture and transport and risks of flooding.

It is important to note that, outside the key regions highlighted here, the skill mostly remains indistinguishable from zero. This can be widespread in particular cases, such as for wind speeds and irradiance in summer or temperatures in winter.

We have only considered the skill of the direct model output from the GloSea5 hindcast here. This represents a minimum level of forecast skill, in two senses. First, the operational forecast ensemble is larger than that available in this hindcast, and it is well established that the forecast skill in GloSea5 increases with the size of the ensemble (e.g., Scaife et al. 2014; Li et al. 2016; Dunstone et al. 2016; Lu et al. 2017).

Second, statistical models linking larger-scale drivers directly to the impact variable of interest may offer further improvement in predictability (e.g., Scaife 2016, chapter 9). This technique has been used for seasonal forecasts in the United Kingdom (e.g., Svensson et al. 2015; Palin et al. 2016; Clark et al. 2017) and is often used already in China (e.g., Xiao et al. 2012; Wang et al. 2013; Peng et al. 2014; Wang et al. 2015; Xing et al. 2016). Research is ongoing to understand the predictability of larger-scale drivers in GloSea5 and how they can be used to improve sector-specific forecasts.

Furthermore, the most user-relevant services are likely to be forecasts of the particular impact of interest to the user, such as energy supply or demand. A next step in developing seasonal climate services on the basis of these results should therefore be to assess the skill of GloSea5 against such direct impacts data, where available from a potential user. The way that forecasts are communicated and handled also affects the usefulness of the forecast (e.g., Taylor et al. 2015; Davis et al. 2016): user engagement is therefore key to optimizing a climate-prediction service. Nevertheless, if co-developed with users and communicated carefully, our results show some areas of very promising skill, allowing the development of improved, skillful seasonal climate forecasting services for specific parts of the energy sector and other sectors in China.

Acknowledgments. This work and its contributors (PB, JL, AS, CH and NG) were supported by the U.K.–China Research and Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. Author CL was supported by the National Natural Science Foundation of China (Grant 41320104007). Author PB thanks Hongli Ren, Jo Camp, Robin Clark, and Margaret Gordon for helpful discussions.

APPENDIX A

Reliability and Sharpness Diagrams and Brier Skill Scores

Model reliability is a description of how closely the forecast probabilities of an event correspond to the frequency of that event being observed in historical data, assessing a conditional bias in the forecast system: for example, we might find that every time the event is forecast to occur with 70% probability it actually occurs only 60% of the time. Once such discrepancies have been characterized, they can then be removed through calibration, resulting in improved forecasts. A full description can be found in Wilks (2011, chapter 8) but we describe the key points for interpreting our plots here.

The reliability diagram for a given class of event is a plot of the observed frequency of the event at times when it was forecast to occur with a given probability. As described in section 2b, we use bins of probability of width 0.1 and pool the event counts from all grid cells in the chosen region of interest. For the set of years in which the event was forecast to occur with a given probability, we plot on the vertical axis the fraction of those years in which the event actually occurred, and then we join the points from each bin with a line.

We mark additional lines in our reliability diagrams (sometimes called an attributes diagram; Hsu and Murphy 1986). The black, solid 1:1 line marks “perfect reliability,” differentiating between underconfident and overconfident forecasts—these will have steeper or shallower reliability lines, respectively, than the “perfect” case. The climatological frequency (e.g., 1/3 for terciles) is marked with a dotted horizontal “no resolution” line: if the reliability line coincides with this line, then it cannot resolve different events into different probabilities, because all forecasts occur at the climatological rate. A dashed line midway between perfect reliability and the no-resolution line is called the
‘no skill’ line, because only points above this line make a positive contribution to the BSS. We shade this region of skill in green.

The Brier skill score measures how much better the forecast system is relative to choosing the climatological rate of occurrence for the event in question (climatology) and can be written as

\[
BSS = \frac{RES - REL}{UNC}. \tag{A1}
\]

Here, the resolution (RES) is the weighted mean square distance between the points and the no-resolution line, the reliability (REL) is the weighted mean square distance between the points and the perfect-reliability line, and the uncertainty (UNC) is the product of the observed climatological frequency and its complement (e.g., \(1/3 \times 2/3\)). The skill is positive if \(RES > REL\), that is, if the points in the reliability line are closer to perfect reliability than to the no-resolution line.

Below each reliability diagram, we include a sharpness diagram, which is a histogram of the distribution of forecasts made in each probability bin. If the histogram is flat, then the hindcast has sampled the full range of possible forecast probabilities and is described as sharp. If it is strongly peaked at the climatological frequency for the event, then the system has no sharpness and mostly just predicts the climatological frequency (climatology). Taken together, the sharpness and reliability diagrams provide a complete description of the joint distribution of observed frequencies and forecast probabilities.

**APPENDIX B**

**ROC Diagrams and ROC Skill Scores**

ROC diagrams describe how well the forecast system can distinguish between classes of events occurring and not occurring [again, see Wilks (2011, chapter 8) for a fuller description]. In practice, we construct ROC diagrams and scores using the same event classes and probability bins as were used for the reliability diagrams, counting events and performing weighted sums over contributing grid cells. Four aggregates of the event counts are made for each probability bin \(p\):

1) the number of hits \(N_H(p)\)—that is, the number of times that the event was forecast with probability \(> p\) and was observed to occur,

2) the number of misses \(N_M(p)\)—that is, the number of times that the event was observed but was not forecast with probability \(> p\),

3) the number of false alarms \(N_{FA}(p)\)—that is, the number of times that the event was forecast with probability \(> p\) but was not observed to occur, and

4) the number of correct rejections \(N_{CR}(p)\)—that is, the number of times that the event was not observed to occur and was not forecast with probability \(> p\).

We then calculate the hit rate (HR) and false-alarm rate (FAR), respectively, as

\[
HR(p) = \frac{N_H}{N_H + N_M} \quad \text{and} \quad FAR(p) = \frac{N_{FA}}{N_{FA} + N_{CR}}.
\]

The ROC diagram then is a plot of HR against FAR for a series of probability thresholds. A skillful system has \(HR > FAR\) and therefore an ROC curve in the top left of the diagram; the 1:1 line (dashed in our plots) delineates no skill, as \(HR = FAR\). We therefore use the area \(A_{ROC}\) under the ROC curve as a measure of skill and scale it to produce a skill score that lies between 0 (no skill) and 1 (perfect):

\[
ROCSS = 2A_{ROC} - 1. \tag{B1}
\]

ROC diagrams are insensitive to calibration of the forecast probabilities and therefore complement the reliability diagrams—they assess the potential usefulness of the forecast system after calibration.

**REFERENCES**


Bett, P. E., and H. E. Thornton, 2016: The climatological relationships between wind and solar energy supply


