CHAPTER

1

Introduction: Why Sub-seasonal to Seasonal Prediction (S2S)?

Frédéric Vitart[*], *Andrew W. Robertson*[†]

[*]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom
[†]International Research Institute for Climate and Society (IRI), Columbia University, Palisades, NY, United States

OUTLINE

A rapid evolution is taking place in weather and climate prediction (Shapiro et al., 2010; Bauer et al., 2015). Historically, there has been a clear separation between weather and climate prediction, despite the fact that both use similar numerical tools. *Weather prediction* refers to the prediction of daily weather patterns from a few days up to about 2 weeks in advance, whereas *climate forecasting* refers to the prediction of climate fluctuations averaged over a season and beyond.[1] This time-scale separation between weather and climate prediction has been accompanied by a divide in the weather and climate research communities, for

[1]The terms *forecasting* and *prediction* are used synonymously here; in some cases, *forecasting* is preferred in the context of forecast use and verification, while *prediction* is more general.

https://doi.org/10.1016/B978-0-12-811714-9.00001-2

the historical reasons described later in this chapter. However, a convergence is taking place, spurred by the growing realization that weather and climate take place on a continuum of time and space scales. Coherent phenomena on a range of scales along this continuum lead to predictability on scales from subdaily, to weeks, months, years, decades, and beyond (Hoskins, 2012). The sub-seasonal to seasonal time range (abbreviated as *S2S*), the focus of this book, sits where the weather and climate scales meet and corresponds to predictions beyond 2 weeks, but less than a season. It is also a key time range for *seamless* weather/climate prediction, in which a single model is used to make forecasts all the way from weather scales to seasonal or longer climate scales or, in a more limited interpretation, that the underlying predictability is seamless across timescales, even if pragmatism dictates the use of different models for forecasting at different lead times (Brunet et al., 2010). We note that the term *S2S* has recently been used more broadly to include seasonal forecasts up to 12 months ahead (NAS, 2016).

Objective weather and climate forecasting can be divided into two branches: empirical (or statistical) forecasting and numerical weather/climate prediction with dynamical models. Empirical forecasting has been practiced in one form or other for more than a hundred years, if not thousands of years (e.g., Taub, 2003); it consists of making forecasts based on past experience or, in the modern era, by using observational data of current and past states of the weather or climate to fit (or train) a statistical model. Empirical methods can be simple (e.g., persistence, where the current weather/climate is predicted to persist for a certain period of time) or more sophisticated (e.g., regression models or discriminant analysis). For example, the analog method of weather forecasting involves examining today's state of the atmosphere and finding days in the past with similar weather patterns (analogs). The forecaster then would predict the weather based on these past analog days. Correspondingly, an analog seasonal climate forecast might be based on past years with similar phases of the El Niño—Southern Oscillation (ENSO).

Numerical weather and climate prediction, on the other hand, uses mathematical (dynamical) models of the atmosphere or Earth system to predict the weather or the climate. The mathematical models used for short-range weather prediction or long-range climate prediction are based on the same physical principles and set of equations, called the *primitive equations,* with the primary distinction being that climate models need to include additional components of the climate system, such as the ocean, depending on the forecast lead time being targeted. These equations are used to evolve the density, pressure, and potential temperature scalar fields and the air velocity (wind) vector field of the atmosphere either on a latitude-longitude grid or in spectral space, through time. The effects of subgrid-scale processes, including convection, radiation, and interactions with the underlying surface, are not treated explicitly but instead are parameterized in terms of the resolved-scale variables. Although empirical methods can be used for sub-seasonal to seasonal prediction, this book focuses largely on dynamical numerical prediction models.

We begin this introductory chapter with a brief history of dynamical weather and climate prediction, together with the World Meteorological Organization (WMO) programs that were created to coordinate these activities and that gave birth to S2S. We then introduce sub-seasonal to seasonal forecasting research and practice, from the discovery of S2S predictability sources, improvements in numerical weather prediction (NWP), development of seamless forecasting, and the demand from applications. The chapter concludes with a summary of the structure of this book.

1 HISTORY OF NUMERICAL WEATHER AND CLIMATE FORECASTING

Numerical weather prediction (NWP) had its roots early in the 20th century, when a better understanding of atmospheric physics led to the establishment of the primitive equations of the atmosphere (Abbe, 1901; Bjerknes, 1904). The first numerical weather forecast was attempted in 1922 by the English scientist Lewis Fry Richardson, in a report called "Weather Prediction by Numerical Process"; he performed his study while working as an ambulance driver in World War I. In this publication, he described how small terms in the prognostic fluid dynamical equations governing atmospheric flow could be neglected, and a finite differencing scheme in time and space could be devised, to find numerical prediction solutions. He constructed a 6-hour forecast of pressure at two points in central Europe by hand. It took him about 6 weeks to produce this 6-hour forecast; unfortunately, it falsely predicted a surge in sea level pressure when in reality, the pressure remained about the same. The number of calculations required to perform weather forecasts is so huge that it was only with the advent of digital computers and the development of numerical methods that weather forecasting in real time became possible.

The first successful computer weather forecast was produced in 1950 with the ENIAC (Electronic Numerical Integrator and Computer) digital computer, taking almost 24 hours to make the 24-hour forecast (Charney et al., 1950). From there, Carl-Gustav Rossby produced the first operational weather forecast (i.e., routine predictions for practical use) based on the barotropic equation in September 1954. Numerical weather forecasting began shortly afterward on a regular basis in the United States, and at around the same time in other countries. For many years, weather forecasts were issued only from a single integration of the atmospheric model from the best estimate of the atmospheric initial condition. Following Edward Lorenz's groundbreaking 1963 paper "Deterministic Nonperiodic Flow," published in *Journal of the Atmospheric Sciences,* which showed how small changes in the initial conditions could lead to very different forecasts due to the nonlinearity of the primitive equations, ensemble forecasts started to be produced operationally in the 1990s. Instead of making a single forecast of the most likely weather pattern, a set (or *ensemble*) of forecasts were produced, giving an indication of the range of possible future states of the atmosphere, and thus the uncertainty of the forecast, stemming from imperfect knowledge of the initial conditions and shortcomings in model formulation. Today, ensemble weather forecasts are initialized using large numbers of perturbed initial conditions, and the model output is often presented in the form of probabilities.

Getting the best possible estimate of the atmospheric initial conditions is central to NWP, and advances in forecast skill over the past 10 years have come in roughly equal parts from improving these estimates and developing models (Bauer et al., 2015). Various methods are used to gather observational data for forecast initialization (radiosondes, weather satellites, and commercial aircraft and ship reports). These observations are generally irregularly spaced and contain errors, so they need to be processed to perform quality control and obtain values at locations that are usable by the model's mathematical algorithms; this process is called *data assimilation and objective analysis.* Then the numerical model can predict how the weather will evolve from its initial state as an initial value problem.

Numerical weather forecasting has improved significantly since the 1950s thanks to improved scientific knowledge, huge improvement in computing capacity, and the advent of satellite data. Computing power has increased by about an order of magnitude every 5 years since the 1980s. Data assimilation algorithms employ the forecast model and use the order of 10^7 observations per day to derive initial conditions that are physically consistent (Bauer et al., 2015). Improvements in forecast skill have been objectively and quantitatively assessed against verifying observations. In the range of 3–10 days ahead, skill has increased by about 1 day per decade, so that today's 10-day forecast is as accurate as the 7-day forecast in the early 1980s, as shown in Fig. 1 of Bauer et al. (2015). The predictive skill in the Northern and Southern hemispheres is almost equal today thanks to the effective use of satellite data providing global coverage.

The first seasonal forecasts issued by a government office were empirical, and they were probably those issued by the Indian Meteorological Department (IMD) in the 1880s; they used Himalayan snow cover as a statistical predictor for the summer monsoon. The work of Henry Blanford and Sir Gilbert Walker, both early directors of the IMD in colonial times, were motivated by the devastating droughts and famines in India in the late 19th century, which, it has been argued (Davis, 2000), gave birth to the modern field of tropical meteorology.

The first dynamical climate model was developed in 1956 by Norman Phillips, who developed a mathematical model that realistically depicted monthly and seasonal tropospheric circulation patterns. Following Phillips's work, several groups began working to create general circulation models (GCMs) based on the atmospheric primitive equations on the sphere. This development closely paralleled that of NWP models, but with lower horizontal resolution to enable longer simulations, and with parameterizations strictly conserving mass and energy necessary for studies of seasonal to interannual climate variability and climate change. The first GCM that combined both oceanic and atmospheric processes was developed in the late 1960s at the Geophysical Fluid Dynamics Laboratory at the National Oceanic and Atmospheric Administration (NOAA).

Climate predictability comes from the relatively slow evolution (i.e., taking months and even longer) of the atmospheric lower boundary conditions such as sea surface temperature (SST), sea ice, soil moisture, and snow cover. For instance, SST anomalies associated with El Niño or its opposite, La Niña, can be predicted a few months in advance (Barnston et al., 2012), leading to predictability in the impact of these anomalies on the atmosphere, such as a reduction of tropical storm activity in the Atlantic, or rainfall over many parts of the tropics, associated with El Niño (Gray, 1984; Ropelewski and Halpert, 1987). However, this impact of SST on daily weather is not deterministic, and the resulting predictability in seasonal averaged weather was called "predictability of the second kind" by Lorenz (1975).

This fundamental distinction between weather and seasonal climate prediction led to the introduction of probabilistic concepts—including ensemble prediction—into the latter well before the former. Seasonal climate forecasts are (or should be) issued in terms of changes or shifts in climatological probability distribution of weather parameters such as temperature and precipitation; climate forecasting on seasonal timescales and longer is not about predicting the exact weather several months or years in advance, but rather about predicting future changes in its probability distribution over large averaging time periods (ranging from seasons to multiple decades). The importance of ENSO-related, tropical SST anomalies as boundary forcing on the atmosphere led to the development of two-tier seasonal forecasting

systems, consisting of separate components for (1) predicting the evolution of tropical Pacific SST, and (2) simulating the atmospheric response using ensembles of atmospheric GCMs in multimodel combination. This two-tier approach popularized the paradigm of seasonal forecasting as a boundary value problem, and it was used in real-time seasonal climate forecasting at several centers, including at the International Research Institute for Climate and Society (IRI) between 1998 and 2016 (Mason et al., 1999). Fundamentally, however, all dynamical prediction is an initial value problem for the evolution of the phenomena that have predictability on the relevant timescale (Hoskins, 2012).

As mentioned already, climate and NWP models are both based on the same set of numerical representations of the primitive equations. However, climate models need to include additional components of the Earth system in order to represent sources of climate predictability on longer timescales. These include the ocean, land surface, and cryosphere, as well as atmospheric chemistry (including aerosols, ozone, and greenhouse gases) and a more detailed representation of the stratosphere. The coupling of GCMs of the atmosphere and ocean, typically developed separately by research groups of atmospheric scientists and oceanographers, remains a big challenge for climate modelers because small imbalances in the surface fluxes between the models can lead to large drifts in climate when the models are coupled together.

The time evolution of the other components of the climate system is usually assumed to be too small to have a significant impact on weather forecasts a few days in advance. This is why weather forecasts historically are based on only an atmospheric global or regional circulation model, in which sea-ice and SST fields are simply persisted from the initial conditions, and with other components of the Earth system set at their climatological values (e.g., aerosols). However, this additional complexity of climate models has been offset by greater intricacy in the formulation of initial conditions in weather forecasting. Atmospheric observation and data assimilation are traditionally associated with the weather forecasting community rather than the climate forecasting community because of the key importance of good initialization for weather forecasting, while seasonal climate forecasts largely rely on predictability of the second kind, associated with the evolution of the SST boundary conditions. Another key difference between weather and seasonal forecasting is the resolution of the atmospheric model. Because the integrations are much shorter in weather forecasting than in seasonal forecasting, weather forecasts are usually produced with much finer horizontal and vertical resolution than climate models. The typical resolution of seasonal climate forecasts, such as in the North American Multimodel Ensemble (NMME; Kirtman et al., 2014), and simulations for the Climate Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012) is around 100 km, whereas global weather forecasts are now produced routinely at a resolution of up to 8 km, and short-range forecasts with regional models have a resolution of a few hundred meters.

Seasonal forecasting today is carried out routinely by 12 WMO-designated Global Producing Centers (GPCs), as well as by a consortium of research and operational centers in North America—the NMME—and other nongovernmental centers, including the APEC Climate Centre (APCC) and IRI; typically, the forecasts are issued toward the middle of every month. All these centers now use coupled ocean-atmosphere (one-tier) models, in which the initial conditions of the ocean, land, and (in some cases) sea ice are prescribed; seasonal climate prediction is largely an initial value problem in these models, as opposed to a boundary value problem in the two-tier approach. The key boundary conditions in these coupled

ocean-atmosphere-land-ice models are prescribed greenhouse gas concentrations; these variations are especially important for making hindcasts for past years, which are needed for assessing forecast skill.

Climate prediction is also carried out on longer timescales to make decadal predictions, and especially to make projections of anthropogenic climate change. The IPCC has coordinated successive climate change assessments based on increasingly sophisticated Earth system models, which are GCMs to which further components of the Earth system relevant to longer timescales have been added (e.g., ice sheets).

2 SUB-SEASONAL TO SEASONAL FORECASTING

As mentioned previously, sub-seasonal to seasonal prediction (forecasts from about 2 weeks to a season ahead) addresses the gap between medium-range weather forecasting and seasonal forecasting. According to the WMO definitions (http://www.wmo.int/pages/prog/www/DPS/GDPS-Supplement5-AppI-4.html), the S2S scale corresponds to extended-range weather forecasting (10–30 days), and the first part of long-range forecasting (30 days up to 2 years). These ranges are approximate, and a committee formed by the National Academy of Sciences in the United States recently defined the S2S range as between 2 weeks and 12 months (NAS, 2016). As discussed previously, there are good historical reasons for the split between weather and seasonal climate forecasting: S2S was considered a difficult time range for weather forecasting, being both too long for much memory of the atmospheric initial conditions and too short for SST anomalies to be felt sufficiently strongly, making it difficult to beat persistence and leading to the notion of a gap between the two ranges.

A pioneering sub-seasonal forecast attempt was made by Miyakoda et al. (1983). This paper showed how the pronounced blocking event of 1977, which generated exceptional snowy conditions over Florida, was successfully reproduced in 1-month forecasts produced by a GCM (Fig. 6 in Miyakoda et al., 1983). In addition, Miyakoda et al. (1986) found some marginal skill in eight January 1-month integrations using a 10-day running mean filter applied to the prognoses. The use of 10-day low-pass filtering is significant because it implicitly recognizes the importance of time aggregation, which introduces a climate forecasting element. The report of successful forecasts beyond day 10 triggered a great deal of interest at that time, and many of the world's operational prediction centers experimented with extended-range forecasts (from 10 to 30 days ahead) (Tracton et al., 1989; Owen and Palmer, 1987; Molteni et al., 1986; Déqué and Royer, 1992).

The European Centre for Medium-Range Weather Forecasts (ECMWF) used its operational forecast model to produce a pair of 31-day forecasts starting at 2 consecutive days for every month from April 1985 to January 1989 (Palmer et al., 1990). These experiments generally showed some moderate skill after 10 days (Miyakoda et al., 1986; Déqué and Royer, 1992; Brankovic et al., 1988), particularly when comparing the forecast to climatology. However, a particularly tough test for extended-range forecasting is to beat the skill of persistence forecasts. At ECMWF, the extended-range experiments described in Molteni et al. (1986) failed to produce forecasts after 10 days that were significantly better than

persisting the medium-range operational forecasts. As a consequence, this experiment did not lead to an operational extended-range forecasting system at ECMWF.

Anderson and Van den Dool (1994) added another pessimistic note to this problem, demonstrating that some apparent high-quality forecasts in the extended range that triggered the initial enthusiasm for monthly forecasting could have occurred by chance. Using the extended-range model from the National Centers for Environmental Prediction (NCEP) Dynamical Extended-Range Forecasting (DERF) (Tracton et al., 1989), they found that after 12 days, the model did not produce better forecasts than a no-skill control. These disappointing results reinforced for many years the idea that the sub-seasonal to seasonal timescale was a "predictability desert." However, interest in the S2S time range revived in the last decade thanks to four factors.

2.1 The Discovery of Sources of Sub-seasonal to Seasonal Predictability Associated With Atmosphere, Ocean, and Land Processes

Although they are not yet fully understood, the most important sources to date are the following:

- The Madden-Julian Oscillation (MJO): As the dominant mode of intraseasonal variability of organized convective activity, the MJO has a considerable impact not only in the tropics, but also in the middle and high latitudes. In addition, it is considered a major source of global predictability on the sub-seasonal timescale (e.g., Waliser, 2011).
- Soil moisture: Memory in soil moisture can last several weeks and influence the atmosphere through changes in evaporation and the surface energy budget, affecting sub-seasonal forecasts of air temperature and precipitation over certain regions during certain seasons (e.g., Koster et al., 2010b).
- Snow cover: The radiative and thermal properties of widespread snow cover anomalies have the potential to modulate local and remote climate over monthly to seasonal timescales (e.g., Sobolowski et al., 2010; Lin and Wu, 2011).
- Stratosphere-troposphere interaction: Signals of changes in the polar vortex and the Northern Annular Mode/Arctic Oscillation (NAM/AO) are often seen to propagate downward from the stratosphere, with the anomalous tropospheric flow lasting up to about 2 months (Baldwin et al., 2003).
- Ocean conditions: Anomalies in SST lead to changes in air-sea heat flux and convection that affect atmospheric circulation. Forecasts of tropical intraseasonal variability are found to improve when a coupled model is used (e.g., Woolnough et al., 2007; Fu et al., 2007).

2.2 Improvements in Numerical Weather Forecasting

The skill of medium-range forecasting has improved continuously over the past two decades, due to model improvements and better data and forecast initialization. These improvements have not been limited to the first 2 weeks. In particular, dynamical models have shown remarkable improvements in MJO forecast skill scores in recent years (Fig. 1). About 10 years ago, the forecast skill of the MJO by dynamical models was considerably less than that of empirical models (e.g., Chen and Alpert, 1990; Jones et al., 2000a; Hendon et al., 2000),

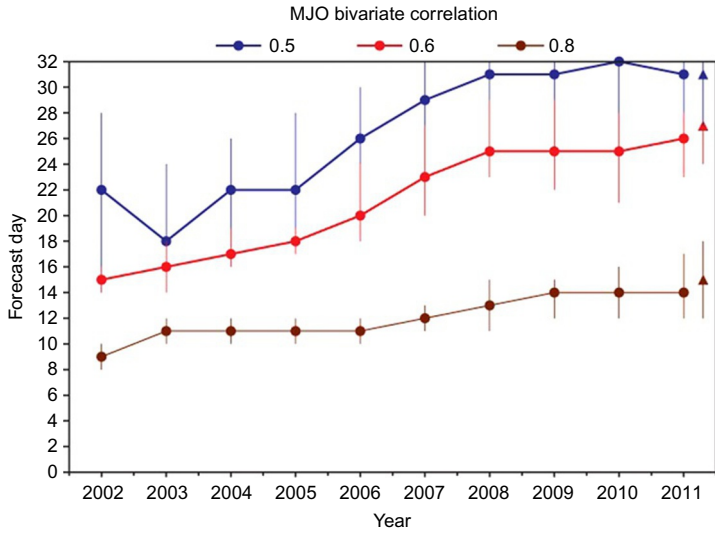

FIG. 1 Evolution of the MJO skill scores (bivariate correlations applied to Wheeler-Hendon index) since 2002. The MJO skill scores have been computed on the ensemble mean of the ECMWF reforecasts produced during a complete year. The blue, red, and brown lines indicate the day when the MJO bivariate correlation reaches 0.5, 0.6 and 0.8, respectively. The triangles show the skill scores obtained when rerunning the 2011 reforecasts with the version of the IFS that was implemented operationally in June 2012 (cycle 38R1). The vertical bars represent the 95% confidence interval (CI) computed using a 10,000-bootstrap resampling procedure (Vitart, 2014).

with skill only up to days 7–10. Recently, skillful MJO forecasts have been reported well beyond 10 days (e.g., Kang and Kim, 2010; Rashid et al., 2011; Vitart and Molteni, 2010; Wang et al., 2014; Vitart, 2014). This progress can be attributed to model improvement (e.g., Bechtold et al., 2008a) and better initial conditions, as well as the availability of historical reforecasts to calibrate the forecasts. Vitart (2014) also reported a significant improvement in 2-m temperature weekly mean prediction in the extratropics for weeks 3 and 4. Newman et al. (2003) found some strong predictability of week 2 and week 3 averages in some regions of the Northern Hemisphere using a statistical linear inverse model (LIM). These improvements in numerical prediction have provided an important stimulus for operational centers to revisit the sub-seasonal to seasonal prediction problem.

2.3 Development of Seamless Prediction

As alluded to already, the unifying concept of weather-climate predictability across multiple timescales has become increasingly prevalent over the last decade, as witnessed by several recent publications (Hurrell et al., 2009; Brunet et al., 2010; Shapiro et al., 2010; Hoskins, 2012). Fig. 1 in Hurrell et al. (2009) illustrates this concept, in which slower larger-scale climate phenomena provide the background for smaller and faster scales, while the integrated effects of the latter can exert important feedback about the former. This concept is epitomized at the S2S scale, which bridges between planetary scale phenomena (including ENSO and the MJO) and local daily weather conditions. Weather and climate have always

been applied sciences, and indeed the quest for better early warning of high-impact weather events has contributed to the revival of interest in S2S. There is a long history of studies of so-called low-frequency variability ($T > 10$ days) of midlatitude weather, beginning with the work of Rossby and his contemporaries on index cycles describing sub-seasonal vacillations between blocked and zonal flows in the Northern Hemisphere midlatitudes. This early work led to studies on multiple equilibria and weather regimes (Charney and DeVore, 1979; Charney and Straus, 1980; Reinhold and Pierrehumbert, 1982) and the application of dynamical systems theory to S2S timescales (Ghil and Robertson, 2002).

In terms of forecasting, atmospheric models are run at the highest-possible resolution to better simulate the representation of weather fronts. Climate forecasting, on the other hand, is based on more complete Earth system models to better represent the evolution of atmospheric boundary conditions with less emphasis on high resolution because the simulation of day-to-day weather variability was not assumed to be fundamental. This difference between climate and weather forecasting is starting to disappear for the reasons summarized in this chapter and pursued in more depth in the other chapters in this book.

How can the theoretical 2-week limit of Lorenz be broken? Two aspects of the seamless prediction paradigm come into play. The first is that the Lorenz limit was derived in the context of midlatitude atmospheric dynamics of baroclinic waves, which have life cycles of about a week. The key to predictability on longer timescales is the existence of predictable phenomena on those timescales, such as the MJO. The second aspect is that averaging on the relevant timescale is critical; while the details of the weather on a specific day will not be predictable beyond 1–2 weeks, weekly or longer aggregates of weather statistics may be predictable in many cases, in the probabilistic sense of climate forecasts. What should the averaging period be for S2S forecasts? Zhu et al. (2014) have suggested that the averaging period should increase in tandem with the lead time, with a 1-week averaging corresponding to a 1-week lead time, and so on, as shown in Fig. 2.

Because weather forecast models are become increasingly skillful and are able to produce skillful forecasts up to 9 days in advance, the weather community has shown increasing interest in using more complex models that include other components of the Earth system to push the limit of predictive skill out a bit more. For example, most operational weather

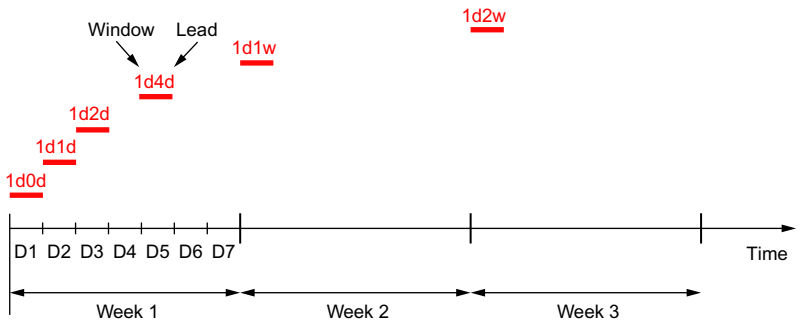

FIG. 2 Schematic of time window and lead time definitions. The horizontal axis represents the forecast time from the initial condition. The expression "1d1d" refers to an averaging window of 1 day at a lead time of 1 day, "2d2d" represents an averaging window of 2 days at a lead time of 2 days, and so on. Note that 1d1d is what is usually called "day 2" in some papers, and "1w1w" is what is usually called "week 2" (Zhu et al., 2014).

forecasting systems still use persisted SSTs (this means that the atmospheric model sees the same SST patterns as in the initial conditions during the full length of the model integration) because it was assumed that SST variations are too slow and small to affect weather forecasting in a significant way. However, ocean-atmosphere interaction has been found more recently to have a significant impact on some atmospheric phenomena, like the MJO (e.g., Woolnough et al., 2007) and tropical cyclone intensity (Bender and Ginis, 2000). As a consequence, some operational weather forecasting systems now include an ocean and a sea ice model, such as at ECMWF (Janssen et al., 2013), which previously had been used only in climate models.

Conversely, there is a growing interest in the climate community to better represent mesoscale weather events in long-range simulation. This is for two main reasons:

- A good representation of weather within climate (synoptic scale events) can feed back into a better representation of the large-scale climate.
- Being able to represent synoptic scale events may allow the direct prediction of impact of climate change on the statistics of the weather. For instance, this would help answer question such as: will global warming impact the number and severity of winter storms over Europe?

There is also a strong interest in testing climate models in weather configurations to help identify systematic errors in models and improve their ability to predict weather events (e.g., a WMO project called Transpose-AMIP, in which climate models are used experimentally for weather prediction). There have also been efforts to run weather forecast models in climate mode to test the evolution of systematic errors associated with slowly varying boundary conditions (e.g., Hazeleger et al., 2010).

From the physical perspective, there are no overriding reasons why weather and climate models should be different, and some operational centers like the United Kingdom's Met Office (UKMO) already use the same atmospheric model for weather and climate forecasting (the unified model, see http://www.metoffice.gov.uk/research/modelling-systems/unified-model). This evolution toward seamless prediction benefits sub-seasonal prediction, in which atmospheric predictability comes from both the initial conditions and boundary conditions. On the other hand, for practical reasons, it may be more efficient to run lower-resolution models with more ensemble members for longer lead times, and the optimal initialization strategy may also depend on the lead time and the phenomena that are the sources of predictability for that lead time.

2.4 Demand From Users for S2S Forecasts

As we have seen, the developments of both weather and climate science and forecasting have strongly use-inspired histories. Thus, it can come as no surprise that societal factors play an important role in answering the question "Why S2S?" The research program of the WMO, a specialized agency of the United Nations (UN) whose mandate covers weather, climate, and water resources with the 191 UN member-states and territories, is organized around two components: the World Climate Research Programme (WCRP), charged with determining the predictability of climate and the effect of humans on climate; and the World Weather Research Programme (WWRP), charged with advancing society's ability to cope with

high-impact weather through research focused on improving the accuracy, lead time, and utilization of weather forecasts. Both WWRP and WCRP have been strong proponents for the development of S2S forecasts for societal benefit, and the WWRP/WCRP S2S Prediction joint research project is the result of those combined forces of improving early warning of climate extremes and understanding human influence on their generation.

While many end-users benefit from applying weather and climate forecasts in their decision-making, many studies suggest that such information is underutilized across a wide range of economic sectors (e.g., Morss et al., 2008b; Rayner et al., 2005; O'Connor et al., 2005; Pielke Jr. and Carbone, 2002; Hansen, 2002). This may be explained partly by the presence of gaps in forecasting capabilities, such as at the sub-seasonal scale of prediction, as well as the large gap between the physical science and end-user domains, which includes the complex task of bringing forecast information into the sphere of multifaceted decision-making.

The sub-seasonal to seasonal scale is especially relevant, as it has the potential to bridge between applications at daily weather timescales and much longer seasonal through decadal climate timescales, where in both cases considerably more societal and economic research has been conducted (e.g., decision and economic valuation studies and climate change impact and adaptation studies). It is therefore an ideal scale to improve forecasts and to evaluate the development, use, and value of predictive information in decision-making. Extending downward from the seasonal scale, a seasonal forecast might inform a crop-planting choice, while sub-monthly forecasts could help inform tactical farming decisions such as when to irrigate a crop or apply fertilizer or pesticides. This would make the cropping calendar a function of the sub-seasonal to seasonal forecast, and thus dynamic in time. In situations where seasonal forecasts are already in use, sub-seasonal ones could be used as updates, such as estimating end-of-season crop yields. Sub-seasonal forecasts may play an especially important role where initial conditions and intraseasonal oscillation yield strong sub-seasonal predictability, while seasonal predictability is weak, such as in the case of the Indian summer monsoon. For example, extending upward from the application of NWP, which is often routine, there is the potential opportunity to extend flood forecasting with rainfall-runoff hydraulic models beyond days to weeks.

In the context of humanitarian aid and disaster preparedness, the Red Cross Red Crescent Climate Centre/IRI have proposed a Ready-Set-Go concept for using forecasts from weather to seasonal (Goddard et al., 2014). In this formulation, seasonal forecasts are used to begin the monitoring of mid- and short-range forecasts, update contingency plans, train volunteers, and enable early warning systems (Ready); sub-monthly forecasts are used to alert volunteers and warn communities (Set); and weather forecasts are used to activate volunteers, distribute instructions to communities, and evacuate areas if needed (Go). This paradigm could be useful in other sectors as well, as a means to frame the contribution of sub-seasonal forecasts to climate service development within the Global Framework Climate Services (GFCS; Vaughan and Dessai, 2014), which provides a worldwide mechanism for coordinated actions to enhance the quality, quantity, and application of climate services; these services aim to equip decision-makers in climate-sensitive sectors with higher-quality information to help them make climate-smart decisions, helping societies better adapt to climate change.

In principle, advanced notification, on the order of 2 to several weeks, of tropical storms, severe heat or cold waves, the onset or uncharacteristic behavior of the monsoonal rains, and other potentially high-impact events, could yield substantial benefits through reductions in

mortality and morbidity and economic efficiencies across a broad range of sectors. Realization of the potential value of such information, however, is a function of several variables, including the sensitivity of an individual, group, enterprise, or organization (or something it values) to particular weather events; the extent and qualities of its exposure to the hazard; its capacity to act to mitigate or manage the impacts such that losses are avoided and benefits are enhanced; and the ability of predictive information to influence its decisions to take action. Unlocking value, therefore, involves much more than creating a new or more accurate prediction, product, or service.

A type of S2S climate forecast has already been popular in applied settings for some time, in the form of *weather-within-climate* seasonal forecasts. For example, the overall frequency of rainy days over the growing season is a key variable for rainfed crops because evenly distributed rainfall is much more beneficial to plants than a few intense downpours with long dry spells in between (Hansen, 2002). It has been shown that in the tropics, seasonal predictability of daily rainfall frequency is often higher at local scale than the seasonal rainfall total (Moron et al., 2007), potentially increasing the usefulness of forecasts by increasing both their salience and credibility (Meinke et al., 2006). Seasonal forecasts tailored to agricultural use thus have started to target the number of rainy days occurring over a particular 3-month season rather than the usual 3-month rainfall average. These weather-within-climate forecasts have seasonal lead time, but the target variable is sub-seasonal. This topic is discussed further in Chapter 3.

3 RECENT NATIONAL AND INTERNATIONAL EFFORTS ON SUB-SEASONAL TO SEASONAL PREDICTION

For the reasons mentioned up to now in this chapter, there recently has been a growing interest in sub-seasonal to seasonal prediction. Ten years ago, only two operational centers, the Japan Meteorological Agency (JMA) and the ECMWF, were producing forecasts at the sub-seasonal time range. Today, at least ten operational centers and most of the WMO GPCs are issuing sub-seasonal to seasonal forecasts routinely.

In 2013, the WWRP and WCRP launched a 5-year joint research initiative, the Sub-seasonal to Seasonal Prediction Project (S2S), with the goal of improving forecast skill and understanding of the sub-seasonal to seasonal timescale, as well as promoting its uptake by operational centers and exploitation by the application communities (Vitart et al., 2012b). A major outcome of this project has been the establishment of a database of near-real-time forecasts (3 weeks behind real time) and reforecasts from 11 operational centers across the world (Vitart et al., 2017): the Australian Bureau of Meteorology (BoM), China Meteorological Administration (CMA), ECMWF, Environment and Climate Change Canada (ECCC), the Institute of Atmospheric Sciences and Climate (ISAC) with the Italian National Research Council, Hydro-meteorological Centre of Russia (HMCR), JMA, Korea Meterological Administration (KMA), Météo-France/Centre National de Recherche Meteorologique (CNRM), National Centers for Environmental Prediction (NCEP) and the UKMO. This database provides an important tool to advance our understanding of the S2S time range and help evaluate the benefit of multimodel sub-seasonal prediction. Sub-seasonal to seasonal prediction is also central to several current U.S. initiatives, such as the NOAA/MAPP initiative

funded by NOAA. There are also efforts in the United States to enhance collaboration between agencies such as the U.S. navy, NOAA, the National Aeronautical and Space Administration (NASA) and the National Science Foundation (NSF) for the development and implementation of an improved Earth system prediction capability (ESPC) on timescales ranging from a few days to weeks, months, seasons, and beyond.

4 STRUCTURE OF THIS BOOK

This book has four parts. Part I, "Setting the Scene," addresses the question of the reasons to use sub-seasonal prediction, which has been briefly discussed in this first chapter. It provides background on NWP and an introduction to ensemble prediction methods. From that basis, it introduces the continuum spatial-scale dependence of the forecast time horizon, with larger spatiotemporal scales predictable for longer into the S2S range and beyond, and discusses the concept of climate predictability of weather statistics based on aggregation in time and space (Chapters 2 and 3). Part I concludes with a theoretical consideration of the potentially predictable modes on S2S scales from the point of view of atmospheric dynamics (Chapter 4).

Part II, the largest part of the book, discusses many of the sources of sub-seasonal predictability identified so far: the MJO (Chapter 5); extratropical waves, oscillations and regimes (Chapter 6); tropical-extratropical teleconnections (Chapter 7); land surface processes (Chapter 8); midlatitude ocean-atmosphere interaction (Chapter 9); sea ice (Chapter 10); and the stratosphere (Chapter 11). Chapter 6 shows an example of how the theoretical framework of dynamical systems provides practical tools for low-order empirical modeling and prediction of S2S variability.

Part III of the book is devoted to several S2S modeling and forecasting issues: the design of forecasting systems used for sub-seasonal prediction (Chapter 12); the generation of ensemble forecasts and data assimilation (Chapter 13); the importance of high-resolution modeling (Chapter 14); the development and testing of S2S forecast products through forecast calibration and multimodel combination (Chapter 15); and verification methods (Chapter 16). A detailed overview of the medium-range and sub-seasonal systems currently used at operational forecasting centers around the world, including their initialization and generation methods, is provided in Chapter 13.

Part IV of the book is dedicated to the use of sub-seasonal forecasts in applications, beginning with the potential to provide early warning of extreme weather events (Chapter 17); seamless prediction of monsoon onset and active/break phases (Chapter 20). This is followed by a chapter on the seamless framework for the early-action use of sub-seasonal forecasts (Ready-Set-Go concept) developed in the humanitarian aid community (Chapter 18); communication and dissemination of forecasts and engaging user communities (Chapter 19); lessons learned from 25 years informing sectoral decisions with probabilistic climate forecasts in the agricultural and energy sectors in Uruguay (Chapter 21); and predicting climate impacts on health at S2S timescales (Chapter 22).

The book concludes with a brief epilogue on prospects for the future of S2S in Chapter 23.

While each chapter is largely self-contained, the references have been consolidated at the end of the book since many are cited in multiple chapters.

CHAPTER

2

Weather Forecasting: What Sets the Forecast Skill Horizon?

Zoltan Toth[*], *Roberto Buizza*[†,‡]

[*]NOAA, Boulder, CO, United States [†]Scuola Superiore Sant'Anna, Pisa, Italy [‡]ECMWF, Reading, United Kingdom

OUTLINE

1 INTRODUCTION

Weather and climate are two aspects of a single reality, the time-evolving atmosphere, as it interacts with the surrounding geospheres. Simplistically, *weather* can be defined as the instantaneous manifestation of this reality, while *climate* refers to weather conditions or their

https://doi.org/10.1016/B978-0-12-811714-9.00002-4

statistics over extended (typically seasonal or longer) time periods. As discussed later in this chapter, the conditions of the atmosphere and its surrounding spheres can be predicted scientifically. In general, more specifics of the expected weather can be foreseen at short lead times, while fewer details of the instantaneous weather are predictable at longer ranges. In particular, specificity about the nature, timing, and position of weather events becomes increasingly elusive as the lead time of forecasts increases.

With advances in the science and technology of prediction, the quality of weather forecasts also has improved, extending the time range for which specific weather forecasts can be made. For example, over the Northern Hemisphere extra tropics, 10-day forecasts of synoptic-scale features are as skillful today as 7-day forecasts were 30 years ago (see Fig. 1).[1] *Sub-seasonal to seasonal (S2S) forecasting* refers to the time range beyond which prediction of weather with finer granularity is lost (today, around 15 days lead time), but lower, sub-seasonal time-frequency and larger spatial-scale variations are still predictable (up to a season or so). After a discussion in Section 2 on the scientific basis for and the evolution of methodologies used in weather forecasting, Section 3 will review, in a historical context, how improved forecast techniques have extended the practical limit of weather forecasting. Forecast techniques used in low-skill

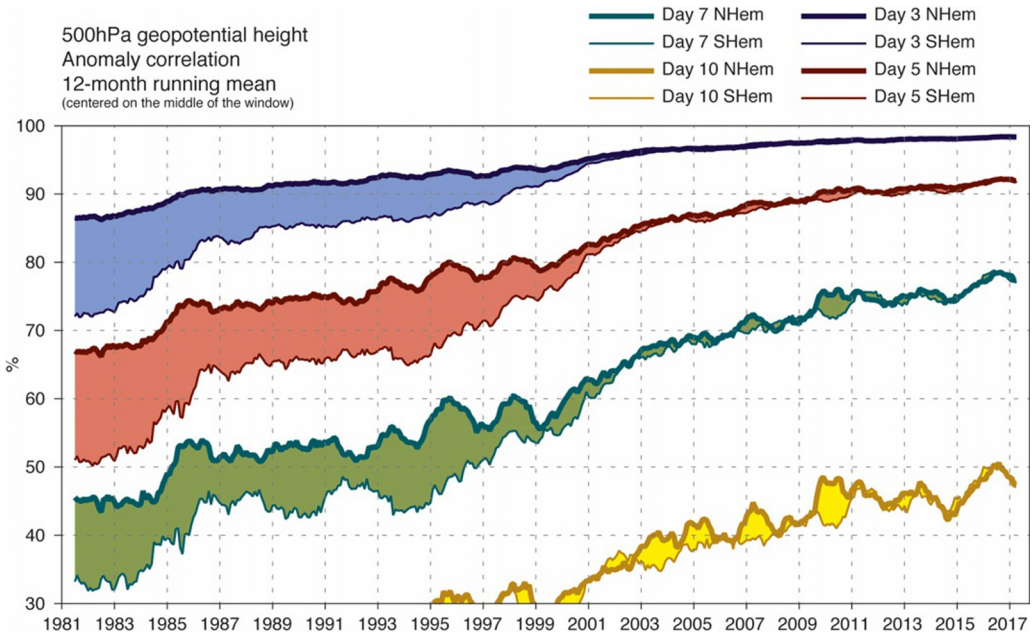

FIG. 1 Monthly averaged forecast skill measured by anomaly correlation coefficient for the 500-hPa geopotential height high-resolution operational forecasts issued by the ECMWF. The pair of *blue, red, green, and yellow lines* show the skill of the 3-, 5-, 7-, and 10-day forecasts over the Northern Hemisphere *(thick lines)* and Southern Hemisphere *(thin lines)*; the shading between the pairs of lines indicates the difference between the skill over the two hemispheres.

[1] Predictability is further explored throughout this chapter, with more formal discussions given in Sections 2 and 6.

environments will also be discussed in Section 4, with a special emphasis on ensemble techniques used so ubiquitously today covered in Section 5. Section 6 reviews how lessons learned from past improvements in weather forecasting may inform S2S efforts to expand the practical limits of predictability and to better exploit forecast skill in the extended range. In particular, we distinguish between *predictability* as conventionally defined related to the spatial and temporal phase of individual weather events (*traceable* predictability), versus predictability of the frequency of such events conditioned on larger-scale (and hence traceable for longer time periods) regimes (*climatic* predictability).

2 THE BASICS OF NUMERICAL WEATHER PREDICTION

The weather that we experience every day depends on atmospheric processes. The atmosphere, of course, is not isolated from, but rather influenced by, its surroundings. Solar insolation, varying primarily on an annual basis, is one of the primary factors driving the general circulation of the atmosphere. Many other slowly varying external factors such as ocean and land surface processes impart an additional level of predictability through their coupling to the atmosphere, which is particularly noticeable on the S2S timescales. Unless noted otherwise, by *predictability*, we refer to current or future scientifically based capacities to skillfully predict the evolution of the atmosphere or its surrounding spheres. Before delving into forecasting the state of coupled systems, however, first we turn our focus to the atmosphere itself. As we will see, some general lessons learned about the predictability of the atmosphere carryover to the more complex coupled ocean, ice, land, and atmosphere systems.

2.1 The Atmosphere as a Dynamical System

Atmospheric processes have been the subject of intense scientific studies. A well-established and critical characteristic of the atmosphere is that its time evolution follows specific rules, and therefore it behaves like a dynamical system. On macroscales, the evolution of the atmosphere is also deterministic, governed by specific physical laws. Importantly, if we know the state of the atmosphere at one point in time, with the use of these natural laws we can predict its state at future times as well. The deterministic nature of the atmosphere (e.g., Richardson, 1922) thus provides the basis for its prediction, which since the 1970s has been done mostly by computers. *Numerical weather prediction (NWP),* as the process is referred to today, will be discussed further later in this chapter.

2.2 Predictability

The behavior of periodic or quasi-periodic deterministic systems such as the solar system can be well predicted for long periods of time relative to the system's characteristic timescale (e.g., a solar year). Under some circumstances, the behavior of periodic or quasi-periodic deterministic systems, however, becomes irregular. This is due to the emergence of instabilities. Forces that previously could balance each other well in a stable fashion become imbalanced, giving rise to a new, dynamically evolving behavior. Interestingly, the atmosphere appears to behave like that: at high levels of viscosity, its laboratory and numerical models follow

regular, periodic, and hence highly predictable behavior that turns aperiodic and much less predictable when viscosity is lowered below a certain level (see, e.g., Ghil et al., 2010 and references therein).

Deterministic dynamical systems with at least one unstable relationship or instability are called chaotic systems. A pendulum suspended via an elastic band (e.g., bungee-jumping) or spring (e.g., Lynch, 2002) is a simple example of a system with aperiodic motions where the centrifugal and gravitational forces temporarily overtake the force of suspension until the spring or band is sufficiently stretched so with its reduced elasticity it can counteract the other two forces. Elasticity in this example is a nonlinear function of the length of a spring or band. Temporally unstable developments in finite size chaotic systems are kept in check by such nonlinear interactions.

If both the governing laws and state of a deterministic system at an instant (such as the atmosphere)[2] are exactly known, the future states, even if the system is chaotic, can be predicted perfectly in perpetuity. The governing laws of real-life systems, of course, are not exactly known. And although the error variance in analysis fields can be estimated, the actual state of natural systems is not known either, due to observational uncertainties. In practice, we can forecast only with imperfect numerical models, from imperfect initial conditions. Errors from the initial condition will amplify in such forecasts due to the instabilities in the atmospheric system, and get convoluted with errors from the use of imperfect models.

Because the true states of natural systems are never known, the actual error patterns in the analysis fields are not known either. The evolution of hypothetical forecast errors, nevertheless, can be explored by studying the evolution of various perturbations to the state of a system. Linear perturbation models and their inverse or adjoint versions are derived by the linearization of the governing equations of a dynamical system (e.g., Errico, 1997). Such models can explore the evolution of infinitesimally small initial perturbations. Linear perturbation studies reveal the innate nature of instabilities of chaotic systems. Without the nonlinear interactions present in the full systems, linear perturbations associated with system instabilities have an exponential, unlimited growth (e.g., Lorenz, 1963; see Fig. 2) that is controlled by a single speed or growth parameter S:

$$v(t) = e^{St} \tag{1}$$

where $v(t)$ is perturbation or error variance at time t. As mentioned earlier, chaotic systems have at least one such instability. Interestingly, in complex dynamical systems, almost any error pattern has a finite projection in all directions in the multidimensional phase space of the system. This includes unstable directions as well. No matter how small, almost any error pattern will project onto instabilities, eventually leading to a complete loss of predictability in chaotic systems.

The exponential growth of linear perturbations clearly indicates that chaotic systems like the atmosphere have only finite predictability. The nature and time after which predictability is lost in realistic systems, however, can be assessed only through the consideration of nonlinear interactions. While nonlinear perturbations are less amenable to theoretical

[2] An estimate of the instantaneous state of the atmosphere, used as an initial condition for NWP forecasting, is called an *analysis*.

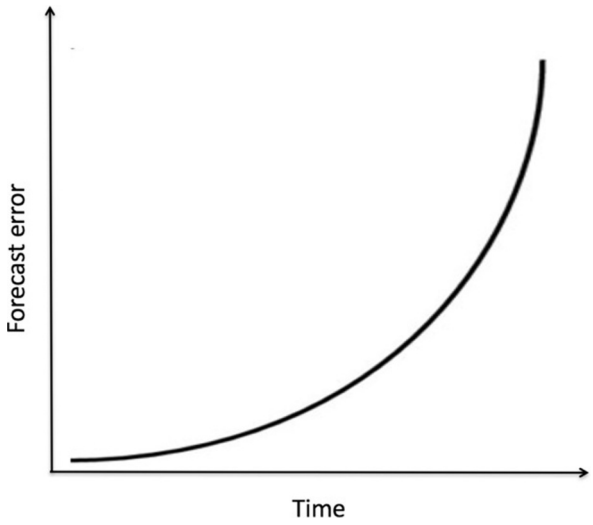

FIG. 2 The exponential growth of linearly evolving perturbation or error variance associated with instabilities in chaotic systems (with arbitrary units).

inquiries, they can be studied by comparing numerical forecasts made with the same model but with slightly different initial conditions (e.g., Yuan et al., 2018). Estimates of error variance between past forecasts and reality, the latter represented by the corresponding verifying analysis fields, also can be used to assess the statistics of error behavior.

Such studies reveal that nonlinear interactions limit the growth of otherwise exponentially amplifying perturbations and errors. Interestingly, in chaotic dynamical systems, the complex, nonlinearly saturating behavior of the initially exponentially growing error variance (v) in time (t) can be well characterized by the following simple and general relationship (Lorenz, 1982):

$$v(t) = \frac{R}{1 + e^{-St}} \tag{2}$$

Notably, the logistic error growth shown in Eq. (2) and displayed in Fig. 3 requires only one additional parameter beyond the speed parameter S that reflects the intensity of the underlying instabilities in linear error growth in Eq. (1). This new parameter is R, the total range or the level of variance at which error growth saturates at long lead times, as all predictability is lost. Note that R, as one of its basic global characteristics, reflects the overall size of the attractor.[3]

2.3 Scale-Dependent Behavior

Atmospheric motions can be documented using different basis functions or phase space coordinate systems. The state of the variables can be described, for example, at selected points in space (i.e., grid depiction), or in special combinations of the gridded variables, such as

[3]The term *attractor* here refers to the collection of all time trajectories that the system can ever visit naturally.

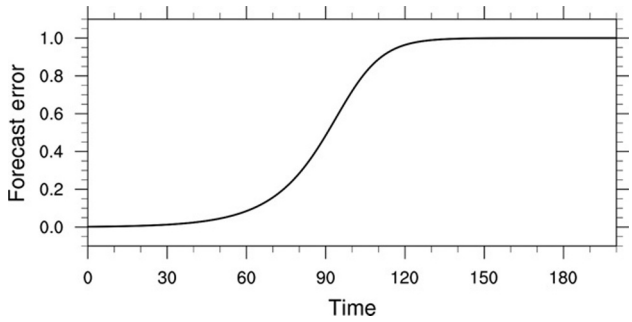

FIG. 3 Logistic growth of nonlinearly evolving normalized perturbation or error variance ($R = 1$, with arbitrary time units) associated with instabilities in chaotic systems. Note the similarity in the initial phase of error growth to exponential growth, as displayed in Fig. 2.

empirical orthogonal functions (EOFs; e.g., Lorenz, 1956), or a coordinate system based on a Fourier decomposition of waves according to their geographical scale (e.g., Orszag, 1969). Using the scale decomposition approach, for example, Rossby (1939) showed that the propagation of atmospheric and ocean features, since called *Rossby waves*, are dynamically dependent on their scale: smaller waves, in general, travel faster.

It turns out that the intensity of instabilities, and hence the speed or rate of error growth (S), also depend on the scale of the motions. In fact, both S and the range (R) parameters in Eq. (2) are a strong function of the spatiotemporal scale of features with which they are associated. We first note that weather is manifested in spatiotemporally coherent features. It follows that the spatial and temporal scales of atmospheric phenomena are connected. Smaller-scale (on the order of hundreds of meters) cumulus clouds, for example, develop as perturbations upon the prevailing environment much faster (in tens of minutes) than large-scale (hundreds of kilometers) extratropical cyclones that evolve over several days. For smaller systems, this results in much faster nonlinear perturbation growth (i.e., larger S) that, simply due to their size, saturates at a low energy level (smaller R).

The same instabilities that force the development of features in the flow are also responsible for the emergence of errors that result in modified, more or less intense, or lack of features. Therefore, faster perturbation growth corresponds with faster error growth, as well as a faster loss in predictability for smaller-scale features. A simple measure of forecast or predictive skill is the correlation between forecast and verifying analysis anomalies taken from the time mean or climatological conditions. An anomaly correlation (AC) or pattern correlation (see, e.g., AMS, 2000) of 1 corresponds to forecasts that perfectly capture the analyzed variability of weather, while zero indicates no skill at all relative to climatological information. As seen in Fig. 4, while features with total wave number 10 (approximately 3000 km horizontal scale, such as a large-scale extratropical cyclone) can be predicted with a useful level of skill (i.e., above 0.6 AC) out to 6 days ahead, smaller-scale features (total wave number 60, about 500 km, e.g., convective clouds organized into mesoscale convective systems) can be skillfully predicted only 1 day in advance. The predictability of individual clouds (not shown in Fig. 4) is even shorter.

AC measures skill in terms of how well forecast and verifying analysis patterns are aligned; therefore, it is is insensitive to the amplitude of errors. Analysis and forecast error

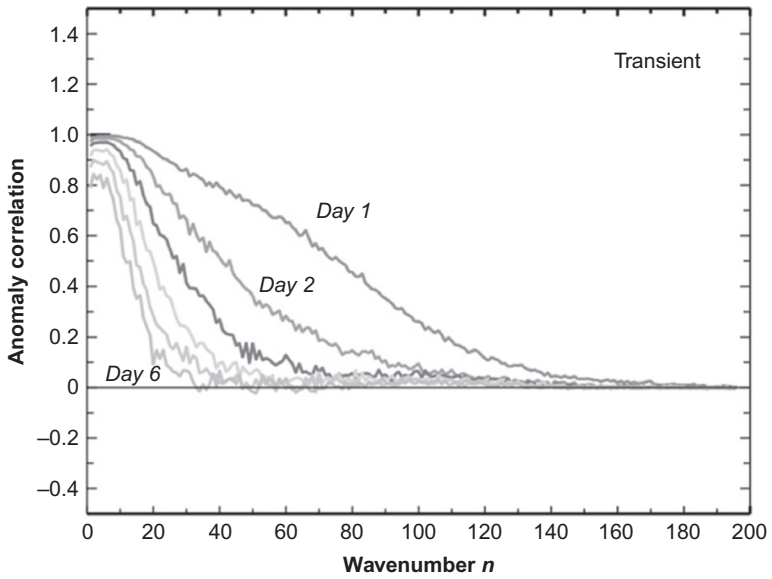

FIG. 4 The AC between Canadian Meteorological Center/Meteorological Service of Canada global 500-hPa forecast anomalies, valid for days 1 (right) through 6 (left), and the verifying analysis anomalies, as a function of total wave number (from the largest scales on the globe to 200 km at $n = 200$), for January 2002. *Adapted from Fig. 6 of Boer, G.J., 2003. Predictability as a function of scale. Atmos. Ocean, 41 (3), 203–215.*

variance,[4] on the other hand, measures the amplitude of errors at each grid point. As the forecast features become uncorrelated, forecast error variance saturates at an R value twice the level of climatological variance for the scales in question (e.g., Peña and Toth, 2014). As seen in Fig. 4, predictability is lost faster on smaller scales. Fig. 5 offers a quantitative assessment of the predictability of the remaining NWP forecasts with progressively longer lead time as a function of total wave number. As noted earlier, forecasts for smaller-scale features (characterized with larger wave numbers) have a smaller climatological variance, and therefore a lower saturation value R (see the heavy solid line on the right side of Fig. 5), corresponding at each scale to the sum of the climatological variance for the model representing nature, and another model used to produce forecasts. The dashed and lighter solid curves in Fig. 5 stand for the error variance in the initial (analysis) and 1–14 day forecast fields, respectively; for synoptic and smaller scales (wave number > 8),[5] the difference between these and the heavy solid curve can be interpreted as the information or predictability remaining in the NWP fields.

[4] After Peña and Toth (2014), analysis and forecast error is interpreted here as the difference between reality interpolated to the NWP grid and the analysis or forecast fields.

[5] Note that due to the reference data used by Privé and Errico (2015) to define anomalies, the thick line in Fig. 5 does not represent an accurate estimate of climatological random error variance for the largest scales.

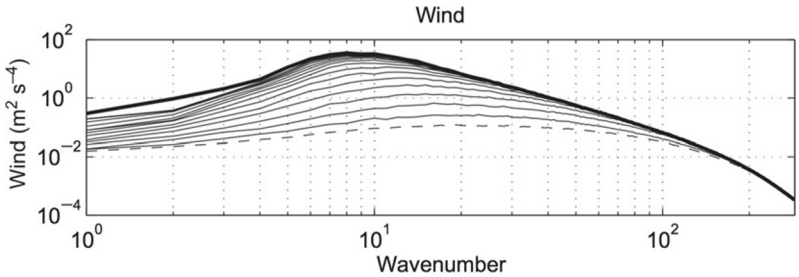

FIG. 5 A 356-hPa rotational wind analysis (dashed line), 1–14-day lead time (*solid line*, from low to high), and random forecast (*R*, *heavy solid line*) error variance as a function of the total wave number for a July–August observing system simulation experiment (OSSE, adapted from Fig. 1 of Privé and Errico, 2015), expressed as anomalies from the respective 2-month means (nature run, analysis, or forecast).

As we move from smaller to larger scales (from right to left in Fig. 5), the analysis error variance curve departs from the random error curve around total wave number 200 about 150 km), indicating that in this OSSE experiment, the analysis retains information about simulated reality down to those scales. Forecasts after 2 days, however, already lose information and predictability on scales finer than wave number 100 (about 300 km), and 4 or 7 days into the forecast, predictability is lost on scales finer than 750 or 1500 km (wave numbers 40 and 20), respectively. As we reach a lead time of 14 days (the shorter end of the S2S time range), only a relatively small amount of information remains on planetary scales greater than 5000 km. Discernable differences between the top thin lines indicate that in this OSSE, some predictability still may be left on the largest scales out to 14 days lead time, even though nature was simulated without full coupling with slowly changing oceanic or ice conditions. If we apply the generic logistic relationship for the description of error growth on progressively larger scales, *S* monotonically decreases, while *R* increases up to, and attains its maximum value at, around about 4000 km (wave number 8).

In summary, while determinism makes predictions possible, chaos, in the presence of initial errors and model imperfections, places limits on the extent of predictability. In particular, with the current level of analysis errors, predictive information on individual events beyond 14 days is restricted to slowly evolving, planetary-scale waves. The position of finer-scale details is predictable only on shorter timescales. The implications of these limitations on S2S forecasting will be further discussed in Section 6.

2.4 Coupled Systems

So far, we have discussed predictability and forecasting in the atmosphere with prescribed boundary conditions, uncoupled from its surrounding spheres. In reality, the atmosphere is in two-way interactions with the ocean, land, and cryosphere. Such two-way coupling with these other systems make the atmospheric circulation more complex. However, the higher level of complexity, due to the rather long "memory" of some high-energy oceanic, ice, and land surface processes, may actually extend the predictability of weather or its statistics in the coupled system. In fact, many of the processes in the coupled system act on a slow timescale compared to those in the atmosphere. A significant portion of the energy of the

coupled system, for example, is locked into annual or decadal scale variations in deep ocean circulation, deep soil moisture and temperature, or sea and land ice fluctuations. When the atmosphere is coupled to these systems, its energy spectrum also shifts to slower scales, giving rise to potential predictability in the atmospheric portion of the full coupled system.[6]

In particular, the coupled ocean-atmosphere system enables the emergence, or enhances the influence of, some slower-evolving instabilities. These in turn led to the formation of longer-lived phenomena such as the El Niño-Southern Oscillation (ENSO). As is the case with the uncoupled atmosphere, the dynamics of the coupled system both supports (due to determinism) and limits (due to chaotic behavior) the predictability of the entire coupled ocean-land-ice-atmosphere system (and, within it, the atmospheric portion of the system). Many of these phenomena, arising from coupling with slower-evolving processes in surrounding spheres and associated with slower-growing instabilities and errors, will be studied in more detail in upcoming chapters.

3 THE EVOLUTION OF NWP TECHNIQUES

Information about future weather originates from observations of its current state. Observations are made either in situ or remotely, from space-, air-, or ground-based platforms. Observing systems typically provide sparse (in space and time) and incomplete (in terms of variables) coverage, as well as inaccurate data about the natural system. To mitigate this situation, data assimilation (DA) must statistically spread information about the state of the system from observed to unobserved sites, times, and variables. Temporal extrapolation of information from past observations is a critical part of the DA process because the first guess for each analysis is provided by a short-range NWP forecast from the previous analysis. In a continual DA-forecast cycle, the DA step statistically combines information from the latest observations with the first guess forecast.

Analyses contain both random (originating from the observations, first guess forecast, and statistical procedures) and growing (originating from the first guess forecast) errors. Growing errors, by definition, amplify in NWP forecasts of chaotic systems, while random errors typically decay. As Toth and Kalnay (1993, 1997) noted, DA-forecast cycles also act as a retaining filter or amplifier of the growing type of error. A major emphasis in NWP development is the reduction of errors in the analysis, especially the growing errors that by definition, influence most the quality of forecasts. Atmospheric analyses are discussed more in Chapter 13.

NWP models are used to approximate the temporal evolution of nature in spatiotemporally discretized fashion on a finite grid. At the core of NWP models are prognostic equations that describe the relationship between the state of the system at two consecutive time steps. Forecasts are produced by the temporal integration of such equations. Natural processes are truncated in space, time, and in the range of represented physical processes. The effect on the resolved scales of some physical processes happening on finer scales unresolved by the

[6]Note that the predictability of the atmosphere can be realistically assessed only in a coupled ocean-land-ice-atmosphere system. If the boundary conditions of the atmosphere are prescribed, the time range of skillful forecasts for the atmosphere will be distinctly different (e.g., Goswami and Shukla, 1991).

models are statistically parameterized (i.e., physical parameterization of convection), while others are ignored (e.g., electrical processes). NWP models are thus imperfect.

Beyond the chaotically amplifying initial errors mentioned previously, forecasts are also affected by errors related to the use of such imperfect models. A smaller part of random (e.g., truncation-related) errors introduced at each time step projects onto growing directions and will amplify as initial errors do. Other model-related errors are due to the model's attractor being displaced from that of nature and manifest as a systematic drift of forecasts started with observed initial conditions, evolving until the model reaches a state close to its attractor.

How have today's NWP systems developed? Next, we offer a brief overview from 1950 onward. For more information, the readers are referred to Lynch (2006, 2008), Deutsche Meteorologische Gesellshaft e.V. (2000), and the papers listed in this chapter.

3.1 Computational Infrastructure

The concept of and scientific underpinning for NWP were created in the 1920s (Richardson, 1922). Due to the large amount of calculations involved, its practical implementation, however, had to wait for the invention of electronic computers in the 1950s. Computational power has continuously increased since, allowing data processing, DA, and forecast applications to become higher in spatial and temporal resolution and ever more sophisticated. Until the 1990s, this increase manifested in ever-faster processing speed, while the number of processors for parallel computing has increased dramatically since.

3.2 Observing Systems

Both surface and upper-air observations (Fig. 6) are indispensable for the initialization of three-dimensional (3D) NWP models. Surface in situ observing networks have been developed since the late 1800s, while the upper-air network was developed during World War ll and in the 1950s and 1960s, partly in support of NWP forecasting. A key development that made NWP possible was the decision to share observations. Thanks to the work of the World Meteorological Organization (WMO, https://www.wmo.int/pages/index_en.html), observation standards were defined, and agreements were put in place to collect and exchange observational data in a timely manner. This led to a further expansion of the global network of observations that allowed NWP centers to generate more accurate estimates of the initial state of the atmosphere, which is required to generate operational weather predictions.

In particular, it is worth discussing two key projects. The first is the World Weather Watch (WWW, http://www.wmo.int/pages/prog/www/index_en.html), established in 1963. Since then, the WWW has been one of the WMO's core programs, combining observing systems, telecommunication facilities, and data-processing and forecasting centers. Today, the WWW, operated by WMO member-states, makes meteorological and related environmental information freely available for the provision of efficient services in all countries. The second project, started in 1967, was the Global Atmospheric Research Program (GARP). GARP, which ran until 1982, helped advance research in the field of weather prediction, including the organization and coordination of several important field experiments, such as the GARP Atlantic Tropical Experiment in 1974 and the Alpine Experiment (ALPEX) in 1982. One key

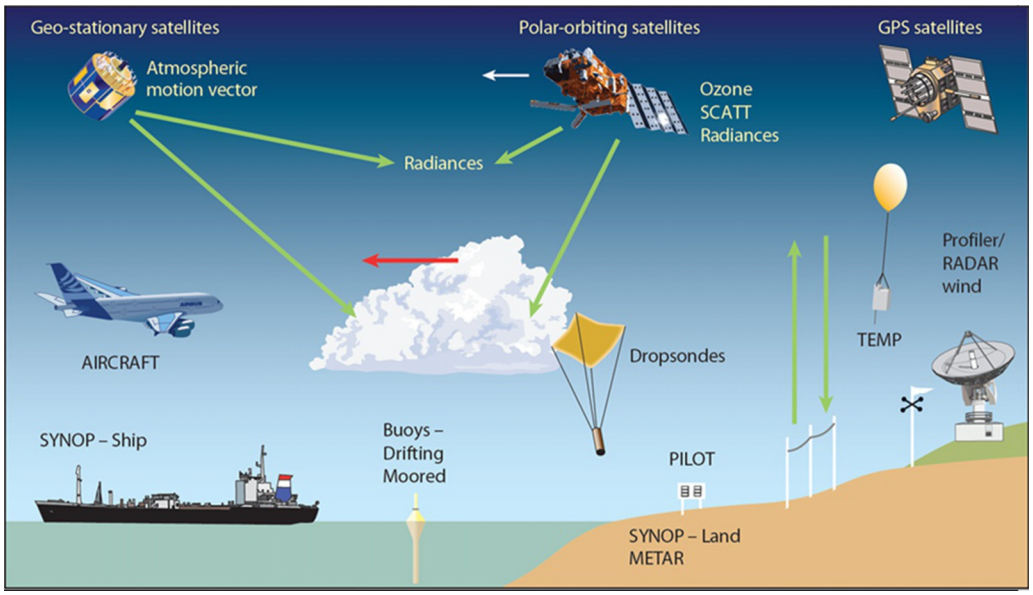

FIG. 6 Schematic of the most common atmospheric in situ and remote observing systems. *From https://www.ecmwf.int/en/research/data-assimilation/observations.*

component of GARP was the First GARP Global Experiment (FGGE), which took place in 1979. FGGE contributed to the extension of the observations of the WWW. These field experiments contributed to the availability of a larger number of, and more accurate, observations, allowing scientists to better test and validate their models.

No credible cost-benefit analysis exists as to the optimal configuration of existing or possibly planned observing systems, so observing system-related developments have been somewhat ad hoc. Although there may be other competitive solutions, governments of developed nations made strategic investments in the design and implementation of satellite platforms. An important milestone of the 1960s was the launch of TIROS-1 (TIROS stands for "Television Infrared Observation Satellite"; see https://science.nasa.gov/missions/tiros), the first weather satellite launched by the National Aeronautics and Space Administration (NASA). TIROS paved the way for the Nimbus program, whose technology and findings are the heritage of most of the Earth-observing satellites that NASA and the National Oceanic and Atmospheric Administration (NOAA) have launched since then. As seen in Fig. 7, the volume of observations in general (including surface, upper-air, and satellite-based measurements) has increased dramatically since the start of NWP in the 1950s. Of particular note is the superexponential growth of satellite-based, mostly radiance observations since the 1980s. Today, about 95% of observations come from instruments onboard satellites. Many of these observations, just as with other remotely sensed data such as lidar and radar observations, contain nonnegligible systematic errors and are indirect (i.e., not of the NWP model variables themselves, but only related to them). Their assimilation, therefore, poses special challenges. The relative role of the observing versus the DA-modeling systems in NWP will be revisited in Section 3.5.

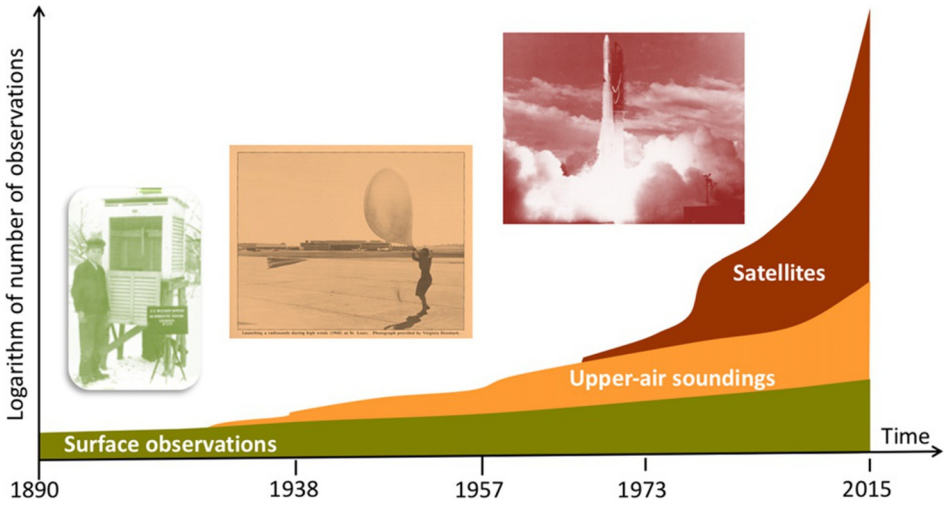

FIG. 7 Schematic illustration of the increasing number of daily surface *(green)*, upper-air *(orange)*, and satellite *(burgundy)* observations available for use in ECMWF reanalysis projects (shown on a logarithmic scale), as a function of time. For reference, in 2015, 44 M of a total of 600 M observations per day were assimilated. *Adapted from Dee, D., 2009. Representation of climate signals in reanalysis. In: Presentation at the Fifth International Symposium on Data Assimilation, Melbourne, Australia, 5–9 October 2009. See at: https://www.dropbox.com/s/ifge2r5wimiyc3h/Dee_2009_Melbourne.pdf?dl=0.*

3.3 Data Assimilation

Perhaps the most critical element in DA is the spatiotemporal propagation of observational information from observations to unobserved analyzed variables. Initial attempts at DA discarded information in all previously taken observations and used persistence or climatology as background fields to fill the large gaps in observational coverage (Bergthorsson and Doos, 1955). Only in the 1960s were NWP forecasts introduced as first guesses in the context of DA-forecast cycles that are still used today. This change led to major forecast improvements, as well as to the birth of modern data assimilation.

For the spatial propagation of information, at each observation site, Gaussian correlation models were used separately in schemes called "optimal interpolation" (Gandin, 1963). Relatively simple relationships served to ensure dynamical balance between the model variables (e.g., normal mode initialization; Errico and Rasch, 1988). A big step forward was the introduction of 3D variational DA (3DVAR). These methods solve a global minimization problem using covariance information across space and variables, derived from a climatology of differences between similar forecast states (e.g., Parrish and Derber, 1992).

It was not until the introduction of four-dimensional (4D) variational DA (4DVAR) that observational information was propagated across variables, space, and time (within an assimilation window) in a flow-dependent manner (e.g., Courtier et al., 1994). This revolutionized the main technique of DA that has not been surpassed since. A plethora of mostly sequential, ensemble-based DA methods have also been developed, but their performance lags that of 4DVAR (see, e.g., Bonavita et al., 2017 and the references therein). Although the basic 4DVAR methods have not been surpassed, various improvements, including ensemble-based

methods to propagate covariance information across assimilation windows, have been developed over the past decade.

A key attribute to the variational approach is that it allows the assimilation of often remotely taken observations of nonmodel variables such as radiation or radar reflectivity. In such cases, observational models (called *observation operators*) connecting the observed and analyzed variables become part of the DA scheme. In this era of proliferating satellite-based radiance and other remotely sensed indirect observations, this feature has invaluable benefits.

3.4 Modeling

Beyond DA enhancements, model development is another key area for improving NWP forecast performance. In 1950, history was made at the University of Pennsylvania by the creation of the first numerical weather prediction using a simplified set of equations on the Electronic Numerical Integrator and Computer (ENIAC). In this effort, a barotropic vorticity model developed at the Massachusetts Institute of Technology (MIT; Charney et al., 1950) was used. Follow-on research in the United States led in 1955 to the start of routine numerical weather prediction under the Joint Numerical Weather Prediction Unit (JNWPU), a joint project between the U.S. Air Force, Navy, and Weather Bureau.

Across the Atlantic, Carl-Gustaf Rossby's group at the Swedish Meteorological and Hydrological Institute used a similar model to produce the first European operational forecast in 1954 (Persson, 2005; Harper et al., 2007). As computers advanced, more complex models capable of simulating a wider range of physical processes became established. Phillips (1956) developed the first primitive-equation numerical model capable of realistically depicting the main features of the troposphere. As a prelude to S2S, the first coupled ocean-atmosphere general circulation model was developed at the NOAA's Geophysical Fluid Dynamics Laboratory (GFDL, https://www.gfdl.noaa.gov/climate-modeling/) in 1960. As scientists developed better models and computers continued to increase in power, more nations started producing operational weather forecasts. In 1966, Germany and the United States began producing operational forecasts based on primitive-equation models. The United Kingdom and Australia followed suit in 1972 and 1977, respectively.

In 1967, a working group of the European Commission suggested that stronger collaboration in natural sciences, and more specifically in weather forecasting, could advance the science and technology in this area. Eventually, this led to the establishment in 1975 of the European Centre for Medium-Range Weather Forecasts (ECMWF, https://www.ecmwf.int/). ECMWF joined the ranks of the already established national meteorological services of the United States, Germany, United Kingdom, Australia, and many other countries, with the aim of tackling the medium-range problem, to assess whether it would be possible to issue skillful forecasts 5–10 days ahead.

Because a simple decrease in grid spacing (i.e., increase in spatial resolution) directly reduces truncation errors, it is not surprising that it leads to the reduction of both analysis and forecast error variance. Changes in horizontal resolution, however, require at least a cubic increase in computer power because a proportional reduction in the length of time steps is also required. Fig. 8 shows, for example, that a single forecast with a 64-km resolution model

implemented at ECMWF in 1991 used a sustained computer power of about 0.001 teraflops (TF). This model configuration had about 3 million grid points. In contrast, the 2017 version of ECMWF's high-resolution global model has about 300 million grid points, spaced 9 km apart, and uses a sustained computer power of about 300 TF.

Numerics such as the rather technical choices for time-step and interpolation schemes, horizontal and vertical grids or other representations, and model variables is aimed at minimizing noise and maximizing the conservation of selected quantities in NWP forecasts. Continued experimentation so far yielded no clear or generally favorable solutions in this area.

Beyond reducing spatial/temporal truncation-related errors, the introduction of new or enhanced physical processes or related parameterization packages is another source of significant gains in the quality of NWP performance. The first numerical models in the 1950s and 1960s had no or very simple physics, while today's models represent an increasingly growing set of interacting physical processes (see Fig. 9).

Coupling an atmospheric model with models of surrounding spheres is another critical path for improved forecast performance. Two-way coupled Earth system modeling can be especially important for the S2S range, as slowly varying processes carry relatively large energy in the ocean, cryosphere, and land surface. Fig. 10 shows the composition of a state-of-the-art Earth system model used for weather and climate studies.

3.5 Improvements in Forecast Performance

Forecast improvements like those shown in Fig. 1 earlier in this chapter are due to the combined effects from all the main components of NWP highlighted here: (1) more and higher-

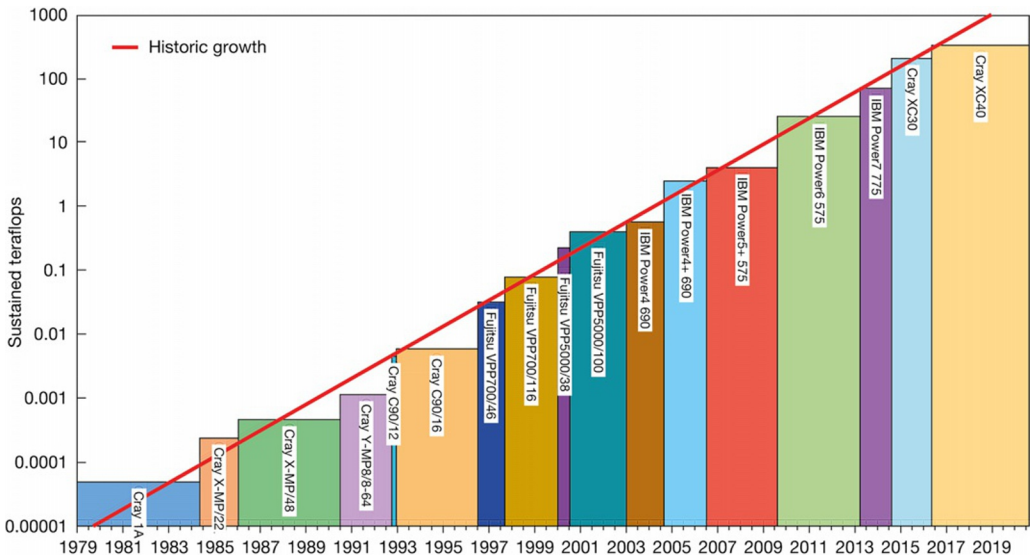

FIG. 8 Time evolution of the computer power installed at ECMWF, expressed in terms of sustained TF (i.e., 10^{12} floating point operations per second) on a logarithmic scale from 1979 to the present.

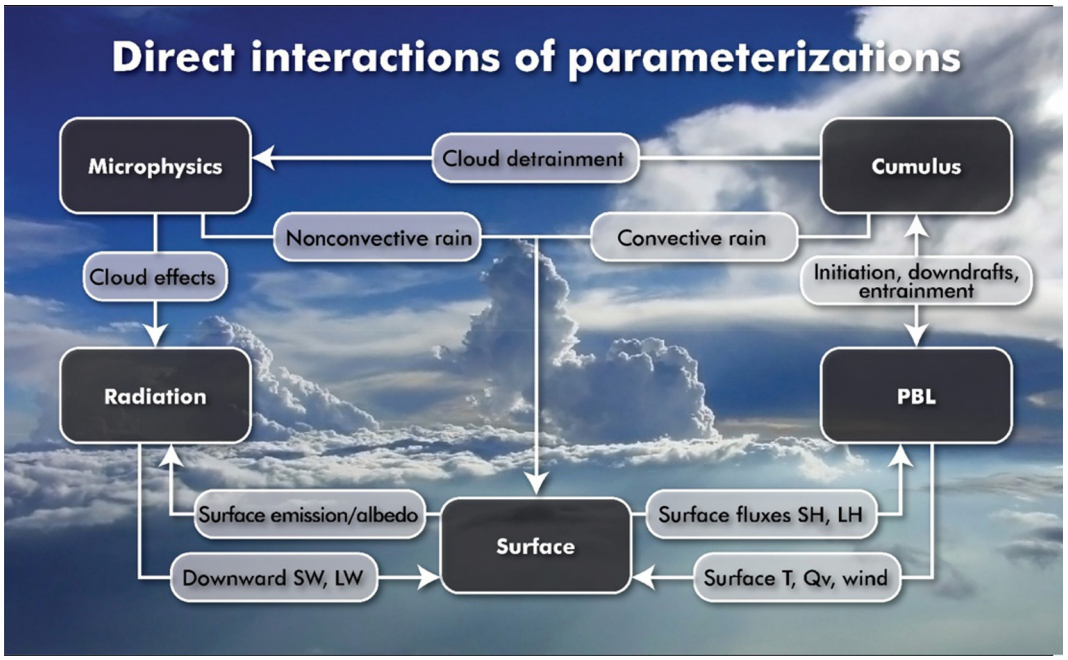

FIG. 9 Interactions *(gray)* between five physical parameterization schemes *(black)* of the Weather Research and Forecasting (WRF) model. Courtesy of the Developmental Testbed Center (DTC).

quality observations, (2) improved DA methods, (3) more realistic numerical models, and (4) more computer power that facilitates many of the advances seen in the other areas. We must note that either by design or necessity, NWP centers often introduce changes in more than one area at the same time, which makes attribution more difficult.

Fig. 11 may shed some light on the relative roles of changes in the observing versus DA-modeling systems on forecast skill. The 15-year period of 1984–98 saw a dramatic increase in the skill of Northern Hemisphere operational forecasts. This period also witnessed a dramatic rise in the availability of satellite observations (cf. Fig. 7). Because the development and maintenance of satellite-based observing systems are arguably the most expensive components of the NWP observing—data assimilation—modeling infrastructure, a naturally arising question is whether the approximately 17-percentage-point gain in operational Northern Hemisphere anomaly correlation scores (shown by the black dashed line in Fig. 11) is primarily due to the costly investments in satellite observing systems, or whether it reflects more the combined effect of less expensive developments in DA and modeling science and computational technology.

Changes in the quality of reforecasts (solid black line in Fig. 11), beyond any natural variability in atmospheric predictability, are only due to observing system changes because the reanalyses and reforecasts were made with a frozen, 1995 National Centers for Environmental Prediction (NCEP) DA-modeling system (Kistler et al., 2001). Remarkably, the Northern Hemisphere reforecasts exhibit no (or only minimal, if any) trend over the 1984–98 period, indicating that DA-modeling improvements dominate the gain in skill, probably at a small

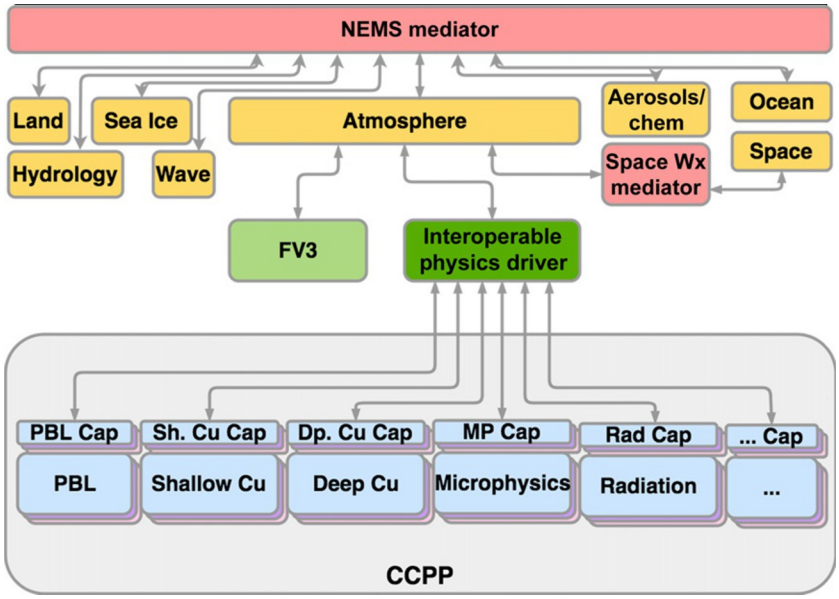

FIG. 10 Schematic diagram of the main models *(yellow)*, their overall mediators *(coral)*, and coupling connections *(lines with arrows)* for the Earth system model of the U.S. Next-Generation Global Prediction System (NGGPS). The dynamics (FV3), physics driver, and physics components of the atmospheric model are shown in light and dark green and blue, respectively. *Courtesy of the Global Model Test Bed of the DTC.*

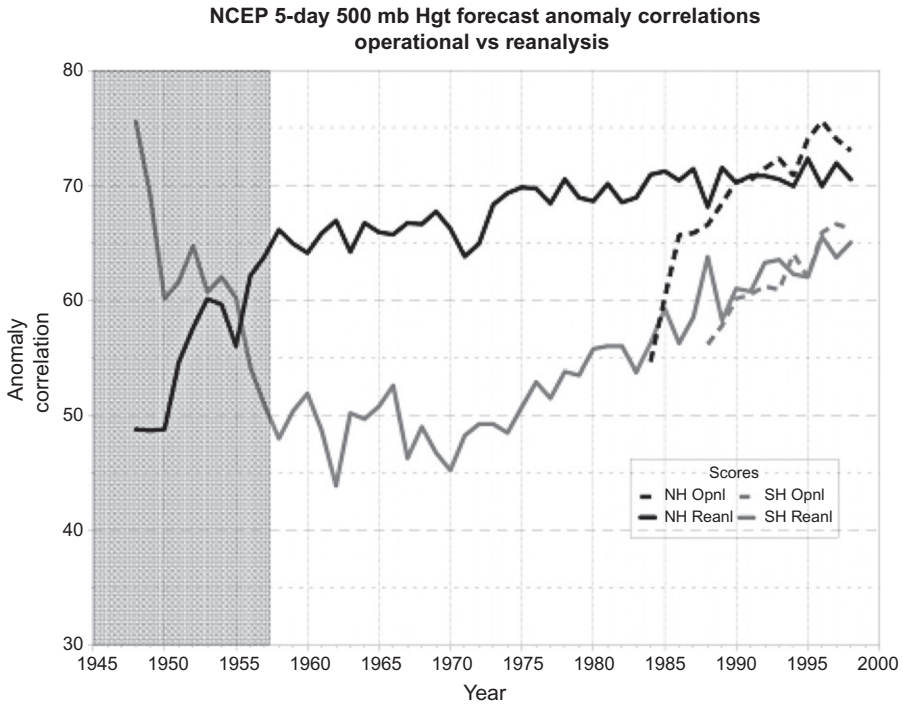

FIG. 11 Annually averaged 5-day anomaly correlation for forecasts from the NCEP-NCAR 50-year reanalysis *(solid lines)* and from NCEP operations—*dashed*, available only after 1984 for the Northern (NH, *black*), and from 1988 for the Southern Hemisphere (SH, *gray*) extratropics. Note that SH scores are artificially high before the 1960s *(shaded)* as verifying analysis fields have too few observations to deviate from the numerical forecasts used as background fields. *Adapted from Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., Fiorino, M., 2001. The NCEP-NCAR 50-year reanalysis: monthly means CD-ROM and documentation. Bull. Am. Meteorol. Soc. 82, 247–268.*

fraction of the cost of concurrent observing system developments. On the other hand, over the in situ, data-sparse Southern Hemisphere, improvements in DA and modeling techniques in the operational forecasts (gray dashed line) appear to bring no added gain beyond what the expansion of the observing network offered in the reforecasts (gray solid line).

Fig. 12 offers further insight into the attribution of skill gains to various changes in DA and modeling techniques. Fig. 12 shows the time evolution of the skill of operational, high-resolution forecasts relative to the skill in a reanalysis-reforecast system that was implemented and "frozen" in 2003 (and never upgraded afterward), thus reflecting 2003 knowledge and technology. Both synoptic-scale (500 hPa geopotential height, 850 hPa temperature, and mean sea-level pressure/MSL), and local weather parameters (2-m temperature, cloud cover, and the 10-m wind speed) are represented in the graph.

The vertical lines in Fig. 12 indicate dates when NWP upgrades were introduced; the red lines show when resolution also changed. It is important to point out that an increase in model resolution is associated with an increase in the resolution used in data assimilation as well, typically allowing more observations to be used. Some highlights from the graph include:

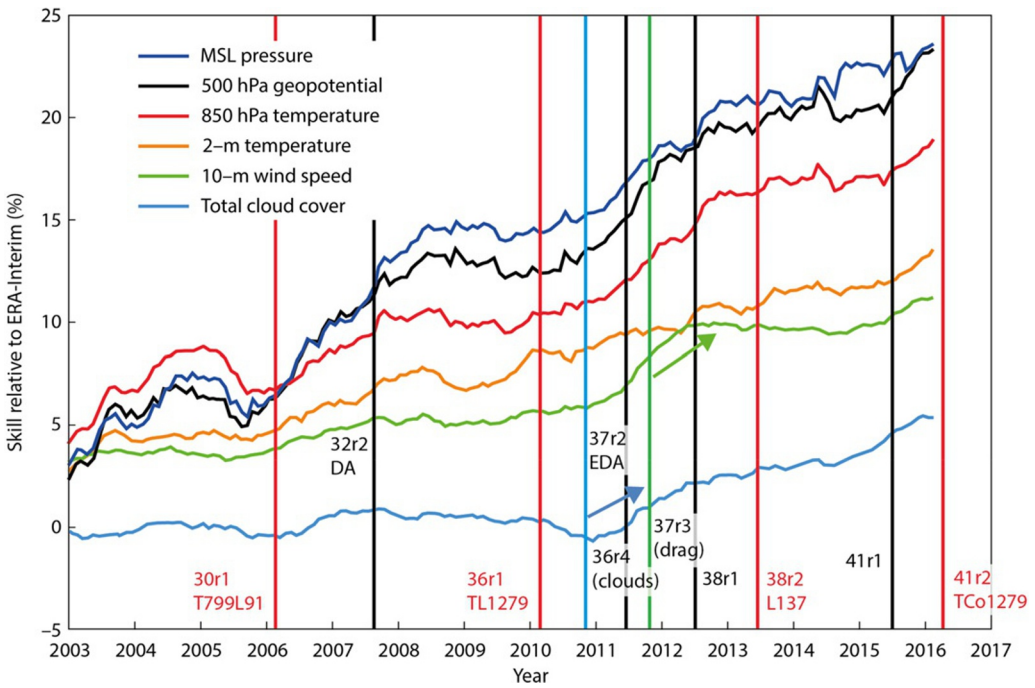

FIG. 12 Time evolution of monthly average skill for operational 5-day high-resolution ECMWF Northern Hemisphere extratropics forecasts from 2003 to 2017, relative to reforecasts made with the frozen ERA-Interim DA-modeling system (Dee et al., 2011). *Blue, black, red, orange, green, and cyan lines* stand for MSL, 500 hPa geopotential height, 850 hPa temperature, 2-m temperature, 10-m wind speed, and total cloud cover. Relative skill is defined as the percentage change of skill in the operational forecasts compared to ERA-Interim reforecasts, both verified against ERA-Interim analyses. Vertical *(red)* lines indicate dates when model (resolution) upgrades were introduced.

- Changes including resolution upgrades (marked 30r1 in 2006, 36r1 in 2010, and 38r2 in 2013) usually bring the largest improvements, but because other model changes are also included, one cannot attribute the improvements only to resolution.
- Upgrades in the data assimilation (e.g., versions 32r2 in 2007 and 37r2 in 2010) also bring substantial improvement.
- Some model changes without an increase in resolution (e.g., 36r4 in 2010 and 37r3 in 2011) can also bring substantial improvement, especially if measured in the variables closely linked with the specific model upgrades (e.g., the upgraded cloud scheme in version 36r4 substantially improved the skill in the prediction of total cloud cover).
- Every change is carefully selected for implementation counts and can lead to improvements.

3.6 Weather Versus Climate Prediction

At the beginning, NWP forecasts were made out to only 24 hours of lead time. With improved techniques, operational weather forecasts were extended out to 3, and then 5–10 and 15 days into the future. These forecasts exploited information primarily about the initial condition of the atmosphere on one hand, and improvements in the modeling of relatively fast weather processes on the other. Ocean, land, and ice processes, if included, were often simplified in NWP models. Meanwhile, the climate community's modeling efforts focused on the initialization and modeling of the slower ocean, land, and ice components of the coupled system. The first successful real-time seasonal predictions of the 1997–98 El Niño event at NCEP (Barnston et al., 1999) gave further impetus for advancing coupled modeling in both the weather and climate communities. With time, it became evident that improved initialization and modeling of the fast and slow components of the coupled system are not exclusive. In fact, both scientific and practical arguments indicated potential gains from a comprehensive and seamless weather-climate forecasting effort, benefiting both weather and climate prediction (Toth et al., 2007). The biggest beneficiary of a converged weather-climate approach, however, has been the intermediary S2S forecasting, where success, without a coupled model initialized well *both* for the fast atmospheric and slow ocean/land/ice components, had been elusive.

The integration of weather and climate forecasting, first started in research, soon reached operations. The full coupling of the earlier atmosphere only models with ocean, sea ice, and land surface models, as well as the significant extension of the NWP forecasts into the weekly range, effectively blurred the line between operational NWP and what originally started as separate monthly or seasonal climate modeling efforts. Today's ECMWF sub-seasonal system, for example, is a seamless extension of its medium-range ensemble, complete with a coupled land, ocean, and atmosphere model out to 46 days twice a week (Vitart et al., 2014a). While the trend is clear, the full integration of forecast effort still has not reached the seasonal time range. For various reasons, including the prevalence of significant biases in fully coupled models, both research and operations in seasonal prediction use somewhat different models and procedures from those applied in weather and S2S forecasting.

As for operational seasonal prediction, it started at NWP and climate forecasting centers in the middle to late 1990s. Since its first implementation, ECMWF has upgraded its seasonal

ensemble prediction system four times (Molteni et al., 2011). The latest version, system-5 (SEAS5), was implemented into operation in November 2017. This system, typical of state-of-the-art efforts, uses a coupled ocean, land, sea ice, and atmosphere model, and it produces forecasts once a month up to 7 months ahead (once a quarter, forecasts are extended up to 13 months). In addition to their weather prediction suites, several national meteorological centers are running seasonal ensembles out to a year. See Chapter 13 for an overview of the global ensembles that are operational today.

4 ENHANCEMENT OF PREDICTABLE SIGNALS

As discussed earlier, predictability is quickly lost on the fine scale, and in a few days even on the moderately large scales. How can we extract a relatively small signal (i.e., the predictable component in a forecast), or alternatively, filter out the large amount of uninformative noise from extended-range weather, S2S, and even longer-range forecasts? Bias-corrected output from an NWP forecast looks just as realistic at a 20-day lead time as at shorter (say, 3-day) lead time. Putting the unlabeled forecast maps side by side, we would not be able to tell which was which. However, we know that the skill of a 20-day forecast is very low at best, whereas the skill of a 3-day forecast is usually high enough for many users to take specific actions. This indeed poses a real challenge for both lay and professional users of weather forecasts, as the possible extraction of any useful information from longer-range forecasts is nontrivial. The presence of spatiotemporally coherent features in the atmosphere, however, offers different ways for the removal of higher-frequency forecast variability that, at a given lead time, may have lost predictability.

4.1 Spatiotemporal Aggregation

Because neither the exact timing nor the location of fine-scale weather features is predictable beyond very short lead times, as an alternative to using the actual prediction valid at a point in time and space, it is a common practice to consider forecast values collected from or aggregated in a spatiotemporal neighborhood of the point of interest (refer to Figs. 4 and 5 for spatial examples, or Roads (1986) for temporal averaging). This is an inexpensive way of generating a range of forecast values, so long as the surrounding terrain is reasonably uniform. One can compute, for example, the spatial and/or temporal mean of forecasts (or their anomalies from long-term means) as a way of removing the unpredictable noise from the forecasts. Alternatively, neighboring forecast values can be used for the inexpensive generation of multiple, a range of, or probabilistic forecasts based on a single NWP prediction (Atger, 2001).

The spatial size or temporal domain of the averaging can be chosen so that a large part of the variance associated with unpredictable scales is filtered out, while most of the predictable signal is retained. Also, 5-, 7- (weekly), 10-, or 30-day (monthly), and district, state, or continental means are various ways that researchers and operational forecasters have attempted to extract and present to users the signal from low-skill forecasts.

It is well understood that due to aliasing effects, box-shaped filters or aggregation areas (with weights of zero outside the zone of interest) used in simple spatial or temporal averaging will introduce some noise into the forecast (see, e.g., Fig. 15-2 in Smith, 2013). Some of the aliasing and noise can be mitigated by using filters with different shapes (e.g., with Gaussian-shaped weights), and by repositioning the filter over each element of interest (i.e., shifting filters or running averages over grid points or time instances).

By combining information from adjacent areas or times, spatiotemporal averaging introduces some errors in a forecast for the selected point of interest. Also, filtering parameters must be chosen to correspond to the actual, lead-time-dependent level of forecast skill. To avoid underfiltering or overfiltering, one must carefully choose filtering parameters to match the actual level of skill.

4.2 Ensemble Averaging

Ensemble forecasting offers a dynamical solution to the abovementioned problems with spatiotemporal averaging. The concept of ensemble forecasting, as discussed further in Section 5 later in this chapter, is rather simple. The analysis of the state of the atmosphere is intentionally degraded by introducing multiple realizations of perturbations that are all within the bounds of the estimated uncertainty in the initial state. In addition to the control forecast from the best analysis, NWP forecasts are made from all perturbed states. The ensemble mean of such forecasts, albeit at very significant computational cost [depending on the number of ensemble members, with O(10) being more costly], offers a flow-dependent filter that, so long as the perturbations are consistent with error statistics, reflects the actual level of forecast skill. An ensemble also offers alternative forecast scenarios valid for each point in time and space, without neighborhood aggregation.

4.3 Removal of Systematic Errors

Numerical models represent, in an approximate way, the dynamics and the physical processes of the real atmosphere. Natural processes are truncated spatially to the model grid space, temporally by the model time steps, and also by the parameterization of some (and ignorance of other) physical processes. Forecast errors thus accumulate due not only to chaotic amplification, but also differences between the dynamics of reality and our models. Some of these differences result in small random errors imparted at each time step of the model into the forecast state, after which it undergoes chaotic amplification just as errors in the initial condition do. Other model-related errors, however, are systematic in nature and manifest as a drift in model forecasts initialized from analyzed conditions. The systematic drift from observed to model-preferred conditions also hinders the use of weather forecasts. A host of statistical methods have been proposed to mitigate this problem. Depending on the level of sophistication and the available of required pairs of forecast-verifying analysis sample data, systematic errors can be estimated over the entire climatological attractor (i.e., by comparing the time means or climatologies of unconditional analysis versus different lead time forecasts), or over different regimes (i.e., via comparing conditional climate means).

Due to the various model truncations mentioned here, many user-relevant atmospheric variables are not part of NWP models. This poses yet another challenge for the use of NWP forecasts. Separate statistical or physical relationships need to be developed between the model prognostic variables and the desired user variables (often on a finer spatial scale than the model's grid) to mitigate the situation. For further discussion of the postprocessing of NWP forecasts, there is a plethora of related literature (see, e.g., Li et al., 2017), and also see Chapter 15 of this volume.

5 ENSEMBLE TECHNIQUES: BRIEF INTRODUCTION

In the past 25 years, NWP has seen a shift from a deterministic approach, based on single numerical integrations, to a probabilistic one, with ensembles of numerical integrations used to estimate the probability distribution function of the future state of the atmospheric variables. This section offers a brief overview of the ensemble approach, which is so ubiquitous now in NWP practice. See Chapter 13 for more details on operational and other near-real-time ensemble forecast systems.

5.1 Background

From the early days of NWP, it was clear that there are some cases when forecast errors remain small even for longer forecast ranges, and others when even shorter range forecasts can have large errors. This operational experience was supported by scientific studies pointing out that due to the chaotic nature of the atmosphere, in the presence of strong instabilities, even small initial errors can grow rapidly and affect forecast quality at shorter ranges.

Yet until the 1990s, the prevailing thinking expressed succinctly by Bengtsson (1991) was that "[w]eather prediction is a well-defined deterministic problem. Starting from a given initial state, any future state can be obtained by integrating the classical Navier-Stokes equations forward in time. Therefore, a weather forecast can, in principle, be calculated in the same way as the motion of the planets or the trajectory of a missile."

However, Bengtsson himself, along with a number of pioneers from the 1950s on (see, e.g., Lewis, 2005) recognized that errors from both the observations and the imperfect models could "contaminate" a forecast. In the 1970s and 1980s, various groups investigated whether case-dependent variations in the quality of forecasts could be estimated in advance (say, when a forecast is issued) to see whether future weather was easier or more difficult to predict than average. In other words, scientists searched for an objective method to associate each forecast with a level of confidence. At the time, various approaches were tested at the major NWP centers. By the early 1990s, a number of groups converged on the idea of using ensemble forecasts. Indeed, 1992 saw the implementation of the first operational ensemble prediction systems at ECMWF and NCEP, following the early work of Lorenz (1975) and others. This development started a new era in NWP prediction, providing multiple scenarios alternative to a single traditional forecast started from the best estimate of the initial state. The implementations at NCEP and ECMWF were to be followed by many others, first at the CMC of the Canadian Meteorological Service in 1995, and then elsewhere.

5.2 Methodology

The main concept behind the ensemble approach is rather simple: generate a set of N perturbed forecasts, each designed to simulate the effect of possible uncertainties associated with the unperturbed (or control) forecast. Then use the N perturbed forecasts to estimate the range of possible outcomes, most probable set of values, and/or the probability that a future parameter (say, temperature at a point in time and space) will be higher or lower than a certain value (Fig. 13).

In the 1980s, different techniques were developed and tested for the generation of reliable and accurate ensembles. One of the approaches tested at the predecessor of the NCEP used lagged forecasts: forecasts initialized recently (say, every 6 hours in the past 2 days) were considered as a lagged ensemble (Ebisuzaki and Kalnay, 1991). The results showed reasonable quality for the medium forecast range, beyond about a week. However, the inclusion of less skillful, older forecasts degraded ensemble quality for the shorter time ranges. Scientists at ECMWF (Hollingsworth, 1980) and NCAR (Errico and Baumhefner, 1987; Tribbia and Baumhefner, 2003) generated ensembles starting all at the same time, but with initial

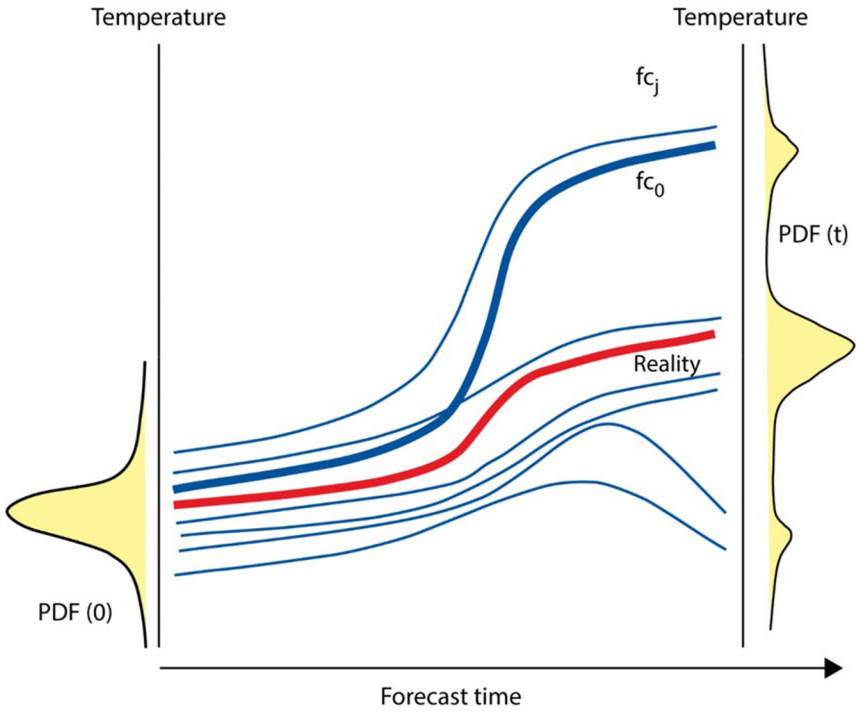

FIG. 13 Schematic of the ensemble approach to the prediction of the time evolution of a selected weather parameter *(red curve)*. Until 1992, NWP forecasts were issued using a single model integration *(bold blue line)* starting from initial conditions with the smallest possible error. Due to initial uncertainties and model approximations, forecasts of a chaotic system diverge from reality. The ensemble approach introduces a set of perturbed forecasts *(thin blue lines)* to estimate a multitude of possible initial and forecast states that in turn can be used to estimate the associated probability distribution functions (PDF, *black curves with yellow shading* at initial (left) and a selected forecast time (right).

conditions perturbed in a random way. This method did not deliver good results either, because the random perturbations did not lead to sufficient forecast diversity: ensemble members remained too similar to provide valuable information on possible future scenarios.

The beginning of the 1990s saw the development and testing of more promising methods both at ECMWF and at NCEP. There are different ways that ensembles can simulate the initial condition and model-related uncertainties. In the first version of the ECMWF global ensemble (Molteni et al., 1996), initial uncertainties were simulated using singular vectors (SVs), which are the perturbations with the fastest growth over a finite time interval (Buizza and Palmer, 1995). SVs remained the only type of initial perturbations used in the ECMWF ensemble until 2008, when perturbations from multiple data assimilation cycles, known as *Ensembles of Data Assimilations (EDAs)*, were also incorporated in addition to SVs (Buizza et al., 2008). Today, SVs remain an essential component of the ECMWF ensemble. They provide dynamically relevant information about initial uncertainties that are linked with forecast errors.

In the first version of NCEP's global ensemble, bred vectors (BVs) were used to simulate initial uncertainties instead of SVs (Toth and Kalnay, 1993). Perturbations in the BV cycle aim to emulate errors in the analysis-forecast cycle (see also the discussion in Section 5). The BV method is based on the notion that due to perturbation dynamics, growing errors have a tendency to accumulate in analysis fields generated by data assimilation (Toth and Kalnay, 1997). Assuming that errors introduced in the assimilation step project in both growing and decaying perturbation directions, one observes that the growing errors amplify, while decaying errors diminish in the pursuant forecast step. Consecutive applications of the analysis-forecast cycle amount to a natural selection (or breeding) of fast-growing errors. The result, confirmed by OSSE studies (e.g., Errico and Prive, 2014), is that growing perturbations dominate analysis error variance compared to a prior expectation of random (neutral) errors.

The ensemble introduced operationally at the Meteorological Service of Canada (MSC) in 1995 adopted a Monte Carlo approach, designed to include as many sources of error as possible. They simulated initial uncertainties due to both observational errors and data assimilation assumptions, as well as, for the first time, model uncertainties (Houtekamer et al., 1996). For the estimation and representation of uncertainties in initial conditions, MSC, similar to Evensen (1994), designed a new data assimilation scheme, a forerunner of a large number of ensemble-based data assimilation plans. Following the Canadian example, a stochastic model perturbation scheme designed to simulate model uncertainties was introduced in the ECMWF ensemble (Buizza et al., 1999).

After the early implementations at NCEP, ECMWF, and MSC, most other operational centers have also introduced ensemble techniques on the global and regional scales, often including schemes to simulate model uncertainties. See Chapter 13 (Ensemble generation: the TIGGE and S2S ensembles) in this volume for a description of the main characteristics of global ensemble forecast systems operational in 2017. These implementations amounted to a paradigm shift in operational NWP from a deterministic approach, based on a single forecast, to a probabilistic one, in which ensembles are used to estimate the probability density function of initial and forecast states.

5.3 Use of Ensembles

Today, it is generally accepted that forecasts must include estimates of uncertainty or confidence that on any day allow forecasters to assess how predictable the future is. Short- and medium-range forecasts, monthly and seasonal forecasts, and even decadal forecasts and climate projections are based on ensembles, providing not only the most likely scenario, but also the uncertainty associated with it. Ensembles of short-range forecasts or data assimilation cycles are also used to estimate the uncertainty in the initial state (analysis).

Probabilistic forecasts, where the probability of the occurrence of a predefined event is predicted, are among the most common ensemble-based products. As ensembles are built around and use the same techniques as other NWP forecasts, it is not surprising that their performance follows similar patterns. Analogous to Fig. 1 (which shows the time evolution of the skill of the ECMWF single, high-resolution forecasts of the 500 hPa geopotential height from 1979 to the present), Fig. 14 shows the time evolution of ensemble-based probabilistic forecasts of the 500 hPa geopotential height over the Northern Hemisphere, from 1995 to the present. Both graphs reveal a steady improvement of forecast accuracy over the years, for example, indicating that a 7-day forecast is as good today as a 5-day forecast was 20 years ago.

Because forecast skill decreases and predictable signals are gradually lost with increasing lead time, a probabilistic approach is paramount. Probabilistic forecasts can be generated statistically based on an ensemble, or even a single unperturbed forecast (e.g., Glahn and Lowry, 1972). Many users, however, are affected by a multitude of weather parameters spread across lead time, space, and variables. Given the sheer number of (and sometimes unforeseen type of)

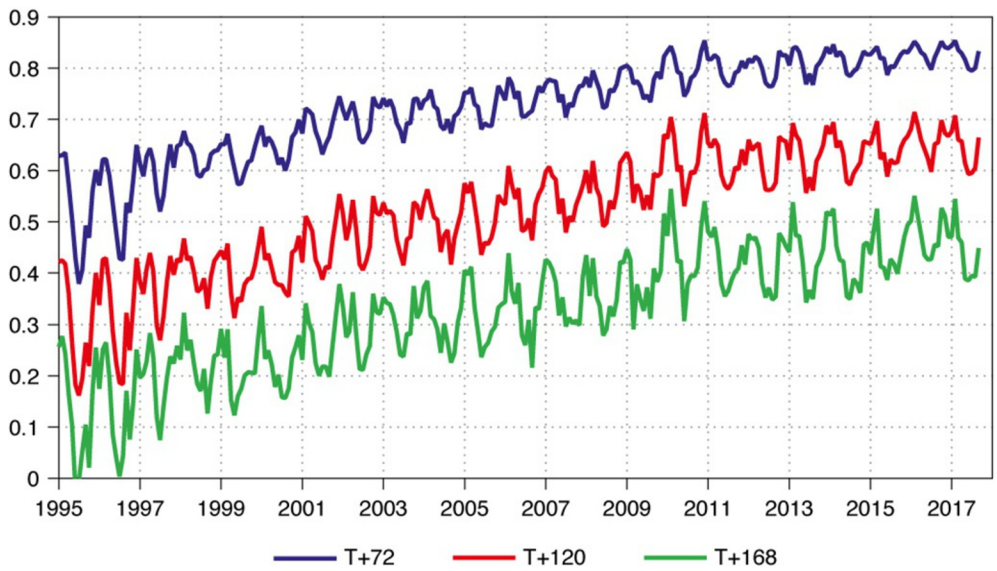

FIG. 14 Time evolution of monthly-average skill using the continuous rank probability skill score (CRPSS) of 3- *(blue)*, 5- *(red)*, and 7-day *(green)* ECMWF probabilistic predictions of the 500-hPa geopotential height over the Northern Hemisphere, measured in comparison with climatologically based probabilistic information.

user applications, the pregeneration of single forecast- or ensemble-based joint or aggregate probabilistic products using statistical calibration methods is simply not plausible. A properly formulated and statistically calibrated ensemble, however, supports innovative applications, including the use of a range of plausible spatiotemporal scenarios, given the initial condition and its uncertainty.

Each realistic ensemble scenario can be fed through user applications such as a model of energy demand from a customer base as a function of various weather and other parameters. Assuming equiprobable sampling by the ensemble members, a rational decision can be made as to optimal, weather forecast–dependent staging for the production and distribution of energy.

Admittedly, many users may lack the sophistication required for the construction of such application models. Nevertheless, an approach like this may offer a quantitative and comprehensive way of capturing case dependent predictability so critical to real-life weather-dependent activities. Ensemble forecasts can also help to prepare for threatening high-impact events in geospheres coupled with the atmosphere that are triggered by weather, where possibly highly nonlinear relationships may limit the use of other, statistically oriented postprocessing methods.

6 EXPANDING THE FORECAST SKILL HORIZON

Recall from the earlier discussion that the term *weather* refers to the time sequence of atmospheric events. As we have seen in Figs. 4 and 5, our ability to pinpoint smaller-scale features in time and space rapidly diminishes as lead time increases. Consequently, at longer lead times, only larger spatial-scale features are traceable in space and time. This confirms Shukla's (1981) early surmise that "the evolution of long waves remains sufficiently predictable at least up to one month." He also suggested that improvements in model resolution and physical parameterizations could extend the predictability of time and space averages even beyond 1 month.

Buizza and Leutbecher (2015) explain that "'forecast skill horizons' beyond 2 weeks are now achievable thanks to major advances in numerical weather prediction. More specifically, they are made possible by the synergies of better and more realistic models, which include more accurate simulation of relevant physical processes (e.g., the coupling of the atmosphere to dynamical ocean and ocean wave models), improved data-assimilation methods that allow a more accurate estimation of the initial conditions, and advances in ensemble techniques." This explanation is consistent with earlier discussions from Shukla (1998) and Hoskins (2013). Shukla (1998) referred to "predictability in the midst of chaos" to explain how skillful long-range predictions of phenomena like El Niño were possible despite fast error-growth rates from small to large scales. Hoskins (2013) wrote about "discriminating between the music and the noise," and introduced the concept of a predictability chain, whereby, for example, "a large anomaly in the winter stratospheric vortex gives some predictive power for the troposphere in the following months."

The reader is referred to Buizza and Leutbecher (2015) for a discussion on the sensitivity of the skill of ensemble-based forecast fields to spatial and temporal filtering, as covered in

Section 4. By applying the same metric to ECMWF ensemble forecasts with increasingly coarser spatial and temporal scales, they showed that forecasts of instantaneous, grid-point fields are skillful up to 16–23 days, while forecasts of large-scale, time-averaged fields have skill up to 23–32 days (because they used ensembles with a maximum forecast length of 32 days, they could not comment on whether the forecasts were skillful for even longer times). These ensemble-based results are consistent with skill estimates for single forecasts, as reviewed in Figs. 4 and 5.

The scale dependency of forecast skill is illustrated in Fig. 15, which shows the forecast time up to which skillful forecasts can be generated (y-axis, logarithmic scale) as a function of the scale of the different phenomena (x-axis, in kilometers). The graph includes not only the ensemble-based estimates from Buizza and Leutbecher (2015), but also other estimates for surface variables such as total precipitation, and for large-scale/low-frequency patterns such as the North Atlantic Oscillation (NAO) and the Madden Julian Oscillation (MJO) and El Niño. The vertical line at 36 km indicates the horizontal resolution of the ECMWF ensemble forecasts in 2013–14, used by Buizza and Leutbecher (2015) to divide the resolved and the unresolved scales. The red lines relative to the instantaneous and finer-scale surface variables are closer to the x-axis, indicating that surface variables are less predictable. By contrast, the blue lines related to teleconnection patterns (e.g., NAO and MJO) and to average Pacific region sea-surface temperature anomalies (SSTA) affected by El Niño are further away from the x-axis and closer to the top-right part of the diagram, illustrating that these large-scale patterns can be skillfully predicted months ahead.

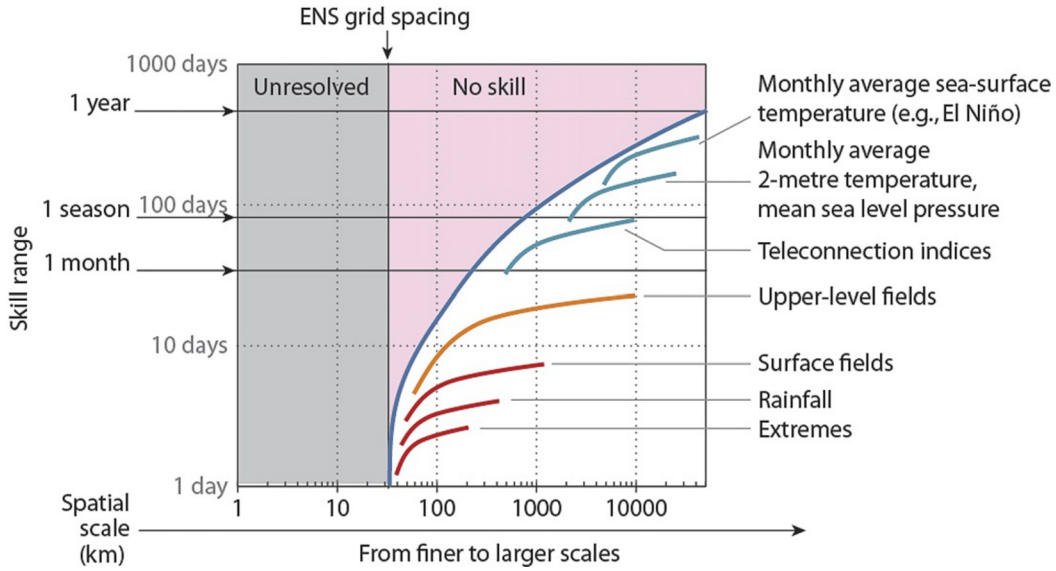

FIG. 15 The Forecast Skill Diagram, which illustrates up to which forecast time ensemble-based, probabilistic forecasts are skillful, is shown (from Buizza et al., 2015). The vertical line labeled "ENS grid spacing" denotes the grid spacing of the ECMWF ensemble used in the Buizza and Leutbecher (2015) study (which was about 36 km), which was used to generate some of the curves in this diagram.

Two further features have been added to the diagram: a blue-line area, drawn schematically to include all the individual lines, and a pink "no-skill: region. The blue line shows the forecast skill horizon for the ECMWF ensemble at the time of the Buizza and Leutbecher (2015) and Buizza et al. (2015) studies. It is still much less than 10 days for very detailed forecasts. The fact that the blue line curves to the right indicates very clearly that the forecast skill horizon, as discussed in earlier sections (cf. the discussion of Boer, 2003), is scale- and variable-dependent. Forecast skill is also a function of geographical location and the season of the year (see Buizza and Leutbecher, 2015). Forecast performance results presented in Fig. 15 reflect the state of the art in science and technology as of the mid-2010s and are a function of errors in the analysis field (i.e., size of initial errors), and errors introduced due to approximations in numerical modeling. Should data assimilation and modeling techniques continue improving, which will be partly due to future S2S research, forecast performance statistics, or practical predictability, will likewise keep improving.

Fig. 15 demonstrates that a portion of large-scale and low-frequency variability can be successfully predicted with today's NWP systems. It is also evident (see the heavy solid line showing an estimate of error variance in random forecasts for rotational wind in Fig. 5) that the energy spectrum of errors, and thus the potential forecast signal, peak around wave number 8 and rapidly drop over the more predictable planetary scales. The predictable forecast signal associated with the planetary scale temperature or other variables of common interest is even lower, while the socioeconomic impact of such slow changes may be only marginal. The introduction of proper coupling of the atmosphere with surrounding spheres in the coming years is expected to raise predictable variance for larger-scale and longer-term motions, but only to a limited extent. It is well understood that high-impact weather events are often triggered by finer scale and much more rapidly evolving features with no traceability beyond a few days whatsoever. Such atmospheric features are often the cause of sudden or catastrophic events in other spheres as well, such as inland flooding, mudslides, snowdrifts, surges, high ocean waves, and the breaking off of ice shelves. In summary, predictability of the more critical small- and moderate-scale motions is quickly lost, while the low level of the remaining predictability in the planetary scales over the extended range is associated with low variability.

Lorenz (1975) and follow-on investigators (e.g., Chu, 1999) distinguish between first and second kinds of predictability, the first influenced by the initial condition of a system itself, while the second by its boundary conditions. Recognizing that in the context of a coupled system, the conditions at the boundary between two subsystems—just as any other variable of a coupled system—are also initial value dependent, we offer a somewhat different perspective on various types of predictability. Following the discussion given so far here, we define the first kind, or *traceable*, predictability as the ability to continuously follow the propagation, emergence, and demise of features in a forecast from the initial time on (see Fig. 16). Fig. 15 assesses this type of predictability in practice, with today's NWP systems. For small-scale features, traceability, as we saw it, is lost at early lead times. Due to the nonlinear interactions between various scales of motion, the statistics (i.e., time frequency) of finer-scale phenomena may be different under distinct larger-scale conditions.[7]

[7]Unlike traceable predictability, this type of behavior is not assessed in Fig. 15.

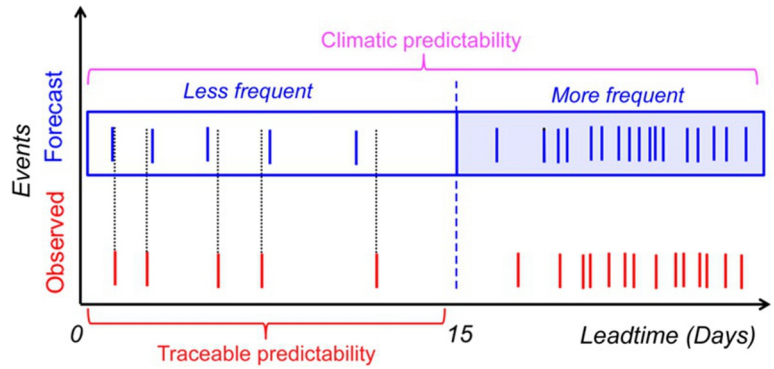

FIG. 16 A schematic illustrating traceable and climatic predictability. On the shorter, weather timescales, individual observed events *(red bars)* are predicted *(blue bars)* with acceptable timing errors (traceable predictability). Low-frequency changes in the frequency or other statistical characteristics of the observed events associated with large-scale regimes, however, can be predicted for longer periods (climatic predictability).

High-impact weather event statistics, for example, are modulated by slowly varying and somewhat predictable phenomena. In other words, high-frequency weather modulations are conditioned on slower varying changes. We refer to the predictability of variations in such frequency and other statistics of finer-scale features, conditioned on the first, traceable kind of predictability of larger-scale motions, as *second kind,* or *climatic predictability.*

The second type of predictability or prediction, therefore, is concerned about climatic frequencies of finer-scale weather phenomena conditioned on the presence of still somewhat predictable large-scale phenomena, as compared to the full climatological frequency of such events. The frequency of tornadoes (see, e.g., the 4–8-day outlooks issued by NCEP's Storm Prediction Center, http://www.spc.noaa.gov/products/outlook/) or hurricanes (see, e.g., seasonal outlooks issued by NCEP's Climate Prediction Center, http://www.cpc.ncep.noaa.gov/products/outlooks/hurricane.shtml) may be below or above overall climatological frequencies, depending on large-scale weather or seasonal regimes that remain somewhat predictable even after the traceability of smaller-scale individual features at any specific place or time is lost. See Chapter 17 of this book for examples of, and more discussion about, the predictability of certain statistics of some of these events. An ensemble of forecasts, for example, may reveal some skillful information on the phase of some still predictable, larger-scale features (i.e., traceable predictability), while beyond their own traceable predictability, the frequency of finer-scale phenomena in ensemble members may reflect frequency statistics consistent with the predicted larger scales (i.e., climatic predictability). This second type of predictability may play a significant role, especially in extended-range S2S predictions.

7 CONCLUDING REMARKS: LESSONS FOR S2S FORECASTING

This chapter reviewed the basis for and limits of weather predictability. We saw how systematic efforts aimed at exploiting the first kind or traceable predictability led the NWP community to the extension of the weather forecast horizon. Numerical modeling and data

assimilation technique development has focused on capturing short timescale behavior primarily in the atmosphere and near its boundaries that are most critical to weather forecasting. Building on these successes, inroads also have been made in predictions beyond the limit of traceable weather forecasting. As in the past decade, NWP adopted an Earth system modeling approach that fully couples atmospheric motions with slower processes, and clear evidence has emerged about the possibility of predicting conditional statistics or the climatology of weather on extended ranges. Aiming for a more thorough exploitation of this second, climatic type of weather predictability, and building on the experience and successes of weather forecasting, S2S must embark on its own systematic path to realistically describe and exploit in both numerical modeling and initial state estimation slowly varying processes in adjoining spheres that in the future can impart significant additional skill in predicting large-scale atmospheric regimes. Statistics of associated high-impact weather can then be derived by ensemble or other methods for more widespread and quantitative socioeconomic applications.

Acknowledgments

The authors acknowledge helpful discussions with Drs. Nikki Prive of NASA, Ligia Bernardet of CIRES at NOAA/GSD, Malaquias Pena of the University of Connecticut, Thomas Auligne of the Joint Center for Satellite Data Assimilation, and Lars Isaksen of ECMWF. Comments by Drs. Shan Sun and Benjamin Green of CIRES at NOAA/GSD, and by the editors, Drs. Andrew Robertson and Frederic Vitart, on earlier versions of the text led to significant improvements in both the presentation and content. Fig. 3 was kindly provided by Dr. Jie Feng, University of Oklahoma.

CHAPTER

3

Weather Within Climate: Sub-seasonal Predictability of Tropical Daily Rainfall Characteristics

Vincent Moron[*], *Andrew W. Robertson*[†], *Lei Wang*[‡]

[*]Aix-Marseille Univ, CNRS, IRD, INRA, Coll France, CEREGE, Aix en Provence, France [†]International Research Institute for Climate and Society (IRI), Columbia University, Palisades, NY, United States [‡]Institute of Atmospheric Sciences, Fudan University, Shangai, People's Republic of China

OUTLINE

1 INTRODUCTION

Sub-seasonal characteristics of tropical rainfall such as rainfall occurrence and monsoon onset date are important for rain-fed agriculture, where long dry spells can ruin a crop and where the onset of the rainy season is a common planting time (Sivakumar, 1988).

https://doi.org/10.1016/B978-0-12-811714-9.00003-6

Sometimes referred to as "weather within climate," it is the daily statistics of weather that ultimately lead to the societal impacts of climate anomalies, such as floods and agricultural droughts. While the timing of individual wet and dry spells cannot be predicted beyond the weather scale, skillful sub-seasonal to seasonal (S2S) forecasts of their frequency of occurrence may be feasible and could be of great societal value.

The seasonal rainfall amount ($\overline{R}$ hereafter, where the overbar denotes time-averaging) is the simplest and most general characteristic of a rainy season because it sums all rainy events during a season. It is generally assumed that $\overline{R}$ is the most predictable precipitation quantity on seasonal time scales at regional scales ($L \sim 100-1000$ km), and seasonal forecasts of $\overline{R}$ (usually on 3-month periods) are currently issued by many forecasting centers on a regular basis (Goddard et al., 2001; Gong et al., 2003; Barnston et al., 2010; Kirtman et al., 2014; Tompkins et al., 2017). Seasonal rainfall predictability is primarily associated with sea surface temperature (SST) anomalies and coupled ocean-atmosphere modes of variation (primarily El Niño Southern Oscillation [ENSO]), but anomalies in soil moisture (The GLACE Team et al., 2004; Douville and Chauvin, 2000), snow cover and sea ice (Cohen and Entekhabi, 1999), and stratosphere-troposphere interactions (Thompson et al., 2002; Cohen et al., 2010) also contribute. The persistent and large-scale atmospheric responses to these anomalous forcings lead to near-homogeneous anomalies of $\overline{R}$ at regional scales. This link between predictability and spatial scale of $\overline{R}$ anomalies comes from the systematic repetition of a near-constant forcing and response across the season, enabling the emergence of a predictable "signal" above the unpredictable "noise." In this context, noise is a statistical quantity and may be seen as the impacts of all atmospheric weather motions that are canceled out through the temporal summation across a season. This signal-to-noise ratio concept is analogous to the familiar ensemble approach used in general circulation model (GCM) simulations forced by boundary conditions (Rowell, 1998). The mean across ensemble members (stations or grid points, respectively) isolates the signal through repetition of the same dynamical response to the forcing(s) in each ensemble member (stations or grid points), independent of its different initial conditions (e.g., various locations), while the spread among the members (stations or grid points) measures the noise; that is, the fraction of the response not determined by the forcing (Shukla, 1998). Thus, metrics of potential predictability based on spatial coherence of observed anomalies, such as the correlogram, decorrelation distance, number of degrees of freedom, etc., are based on a similar concept to potential predictability revealed by GCM ensembles. Note that we do not attempt to answer the downscaling question of how the signal may be locally modified within a near-homogeneous region.

A simple yet instructive decomposition of $\overline{R}$ involves the product of the number of wet days (N_R) (e.g., ≥ 1 mm), and the mean rainfall intensity ($\bar{I} = \frac{\overline{R}}{N_R}$) on wet days. $\bar{I}$ reflects both the instantaneous rain rate (Le Barbé et al., 2002), which is high for convective rainfall, and the duration of rainy events (Ricciardulli and Sardeshmukh, 2002; Smith et al., 2005; Dai et al., 2009), which is related to the spatial scale and the movement of the rain-bearing systems; these range from localized thunderstorms to larger organized systems such as tropical cyclones and mesoscale convective complexes lasting from several hours to a day or more at a fixed location. Ricciardulli and Sardeshmukh (2002) have shown that tropical wet events estimated from satellite images last in mean 4.9 hours over the continents and 6.2 hours over the oceans. Individual thunderstorms can produce very high local rainfall rates (Dai et al.,

2009; Trenberth et al., 2017), which make the interannual variations of $\bar{I}$ noisy even after averaging over a season (i.e., Moron et al., 2007). By contrast, the rainfall occurrence field tends to reflect spatiotemporal hierarchical organization of convection (Orlanski, 1975). Previous analyses of rain gauges (or 0.25 degrees grid points) in several tropical regions, including India and tropical Africa (Moron et al., 2006, 2007, 2009b, 2017), found that the spatially coherent interannual covariations of $\overline{R}$ are mostly conveyed by N_R, while $\bar{I}$ is a far noisier characteristic; the spatial autocorrelation of $\bar{I}$ decays more quickly than for either $\overline{R}$ or N_R due to its dependency on the wettest few days of each season. A similar finding has been obtained with finer space-time data that allow the number and intensity of subdaily wet events to be quantified, finding that interannual-to-decadal variations of precipitation in Sahelian-Sudanian and Guinean Africa are mostly related to changes in the number of wet events, primarily due here to mesoscale convective systems (MCSs) (Le Barbé et al., 2002; Lebel and Ali, 2009). This suggests that the predictability of interannual variations of seasonal amounts at regional-scale over these regions of Africa is likely to stem mostly from changes in the frequency of the MCSs rather than from changes in their intensity.

The basic decomposition of $\overline{R}$ in the previous paragraph could be refined by considering onset and end dates of the rainy season, enabling differentiation, for example, between a dry spell associated with a delayed onset from those occurring within the monsoon period itself (Moron et al., 2015a). $\overline{R}$, then, may be more fully decomposed in terms of the daily rainfall statistics (e.g., rainfall frequency and mean intensity) during each temporal phase of the monsoon season separately. The sources of predictability for these daily rainfall characteristics involves modulation of the various rain-bearing systems (from the individual thunderstorms to MCSs and tropical depressions) by the slow phenomena mentioned here. For example, regional-scale monsoon onset over the Maritime Continent in September-November is almost systematically delayed (advanced) during warm (cold) ENSO events (Haylock and McBride, 2001; Moron et al., 2009a), while interannual variations in rainfall during the core of monsoon season, around December to February, are less spatially coherent and less potentially predictable (Moron et al., 2010). Another shorter source of predictability is provided by convectively coupled equatorial waves (CCEW; Wheeler and Kiladis, 1999; Lubis and Jacobi, 2015), including the Madden Julian Oscillation (MJO; Waliser et al., 2003; Zhang, 2005).

For sub-seasonal forecasts, the time aggregation period is generally only 1 or 2 weeks (Zhu et al., 2014), so a stronger signal or reduced noise will be required in order to obtain a comparable signal-to-noise ratio to the seasonal case. There is increasing evidence that the popular 2-week "weeks 3 + 4" (i.e., 15–28 forecast days after the starting dates) sub-monthly range represents such an opportunity, at least in some cases. Certain constellations of the MJO and ENSO may give rise to windows of forecast opportunity where the signal is sufficiently enhanced (Li and Robertson, 2015). Summing over 2 weeks reduces the weather noise for certain stages of the monsoon's seasonal evolution and emphasizes any intraseasonal mode of variation that is not canceled out over 2 consecutive weeks. Even if the impact of a particular MJO phase over a given region may typically last only less than a week, its impact can still be appreciable over a 2-week period because the opposite phase will not be included in the same 2 weeks.

This chapter presents an analysis of tropical rainfall weather-within-climate predictability based on estimates of spatial coherence calculated from gridded observed rainfall datasets. Running 15-day time windows are used to identify the sub-seasonal modulation. With these

estimates of potential predictability in hand, we harness them to interpret the patterns of anomaly correlation skill seen in European Center for Medium-Range Weather Forecasts (ECMWF) week 3 + 4 hindcasts.

2 DATA AND METHODS

2.1 Daily Rainfall and OLR

Two primary rainfall datasets are investigated. The first is the Indian Meteorological Department (IMD) high-resolution (0.25 degrees × 0.25 degrees) gridded daily rainfall data for the Apr.-Nov. (extended summer monsoon season) 1901–2014 (Pai et al., 2014; Moron et al., 2017). These gridded data were prepared by spatially interpolating daily rainfall from 6955 Indian stations (with varying data availability periods) using distance-weighted interpolation (Shepard, 1968). The interpolated values were computed as the weighted sum of the station data within a search radius of 1.5 degrees. The scheme was locally modified by including directional effects and barriers (Shepard, 1968). The second is the daily Global Precipitation Climatology Project (GPCP) 1.3 (beta version) dataset (1 degree × 1 degree) from Oct. 1996 to Sep. 2016. Daily precipitation values are estimated from multisatellite observations and are calibrated versus rain gauges at a monthly time scale (Huffman et al., 2001). Some computations are also made using rainfall estimates from the pentad CPC Merged Analysis of Precipitation (CMAP) (2.5 degrees × 2.5 degrees) from Jan. 1979 to Dec. 2015 (Xie and Arkin, 1996). The pentad data are just copied to the daily time scales. Lastly, we use the interpolated daily National Oceanic and Atmospheric Administration Outgoing Longwave Radiation (NOAA OLR) dataset (2.5 degrees × 2.5 degrees) from Jun. 1974 to Dec. 2016 (Liebmann and Smith, 1996).

2.2 S2S Forecasts

Reforecasts of total precipitation are evaluated for the ECMWF Variable-Resolution Ensemble Prediction System monthly forecast system (VarEPS-monthly; Vitart et al., 2008), prepared for the WWRP/WCRP S2S project database (Vitart et al., 2017). The atmospheric component of the ECMWF model (version CY41R1) has 91 vertical levels and a horizontal resolution of TCo639 (16 km) up to day 10 and TCo319 (32 km) after day 10. More model details are given in Vitart et al. (2017). Semi-weekly reforecasts of the ECMWF model over the 20-year (Jun. 1995–May 2014) reforecast period are analyzed, corresponding to real-time forecast start dates every Monday and Thursday from Jul. 2015 to Jun. 2016. The ECMWF reforecasts consist of 1 control and 10 perturbed forecasts on each start date, and the ensemble mean skill is evaluated. The GPCP version 2.1 (Huffman et al., 2009) daily precipitation estimates on a 1 × 1 grid are used for forecast validation. The daily data are averaged from day 15 to day 28 to generate week 3 + 4 time series for both forecasts and observations. The ECMWF total precipitation reforecast is interpolated from 1.5-degrees into 1-degree resolution so it can be compared with the GPCP dataset.

2.3 Method to Estimate the Spatial Coherence

The spatial coherence is empirically estimated using the spatial autocorrelation of the interannual ranks of running 15-day amounts (and frequency of wet days receiving $\geq$1 mm) in a 500-km radius. Considering the ranks instead of the amounts themselves reduces the skewness. Other radii, from 150 to 1000 km, leads to similar spatial and temporal modulations of the spatial coherence. In the same way, the main results are not very sensitive to the data because OLR or CMAP datasets lead to similar results even if their lower resolution (i.e., 2.5 degrees vs 1 degree grid) tends to increase the average spatial autocorrelation. Lastly, the spatial autocorrelation in a 500-km radius introduces a weak negative bias about the spatial coherence, where the spatial autocorrelation pattern is anisotropic as around the Pacific and Atlantic Inter Tropical Convergence Zone (ITCZ) (not shown), but computing the area of spatial autocorrelation above a given threshold as $1/e$ (Ricciardulli and Sardeshmukh, 2002) is far more time-consuming. In summary, our main results about the spatial and temporal modulations of the spatial coherence appear to be rather insensitive to data, horizontal resolution, and empirical estimation.

3 RESULTS

3.1 Daily Rainfall Characteristics of the Indian Summer Monsoon

Fig. 1 gives an example of the diversity of sub-seasonal scenarios for two anomalously dry (1986 and 2002) and two wet (1983 and 1988) monsoon seasons, at two 0.25 degrees grid points near Mumbai (18 degrees 56′N, 72 degrees 50′E) and New Delhi (28 degrees 37′N, 77 degrees 14′E). At Mumbai, all four onsets are quite close to the long-term mean (Jun. 13), while the end is anomalously late in the wet years and early in the dry ones. At New Delhi, it is not possible to define onset and end dates due to erratic rain in 1986 (i.e., only three significant wet spells separated by long dry spells; Fig. 1F), while the 2002 season is shorter and appears shifted later than usual, mostly due to a very long dry spell from late June to late July (Fig. 1H). Both dry seasons at New Delhi illustrate the challenge of defining the true onset of the rainy season as soon as fairly long dry spells occur after (or between) the first wet spells, such as around late June in 2002 in New Delhi (Fig. 1H). This uncertainty in onset definition makes prediction challenging. Note also that both dry years share a very long break in July at New Delhi and Mumbai despite the distance of 1100 km between them. The anomalously wet season in 1983 (Fig. 1B) is related to a longer season than usual in New Delhi, but also to more very wet days. Such behavior is also observed in 1988 (Fig. 1D), except for the end, which occurs close to the climatological long-term mean. Unlike at Mumbai, the wet days at New Delhi are clearly more intense during anomalously wet seasons.

Fig. 1 illustrates that a good (i.e., anomalously wet) monsoon on the subcontinental scale (Sontakke et al., 2008), such as in 1983 or 1988, can have very different sub-seasonal evolutions on the local scale (Moron et al., 2017), and emphasizes the importance of predicting these daily statistics. This example also illustrates the difficulty of calculating daily rainfall statistics strictly between the monsoon onset and end dates because these dates themselves are subject

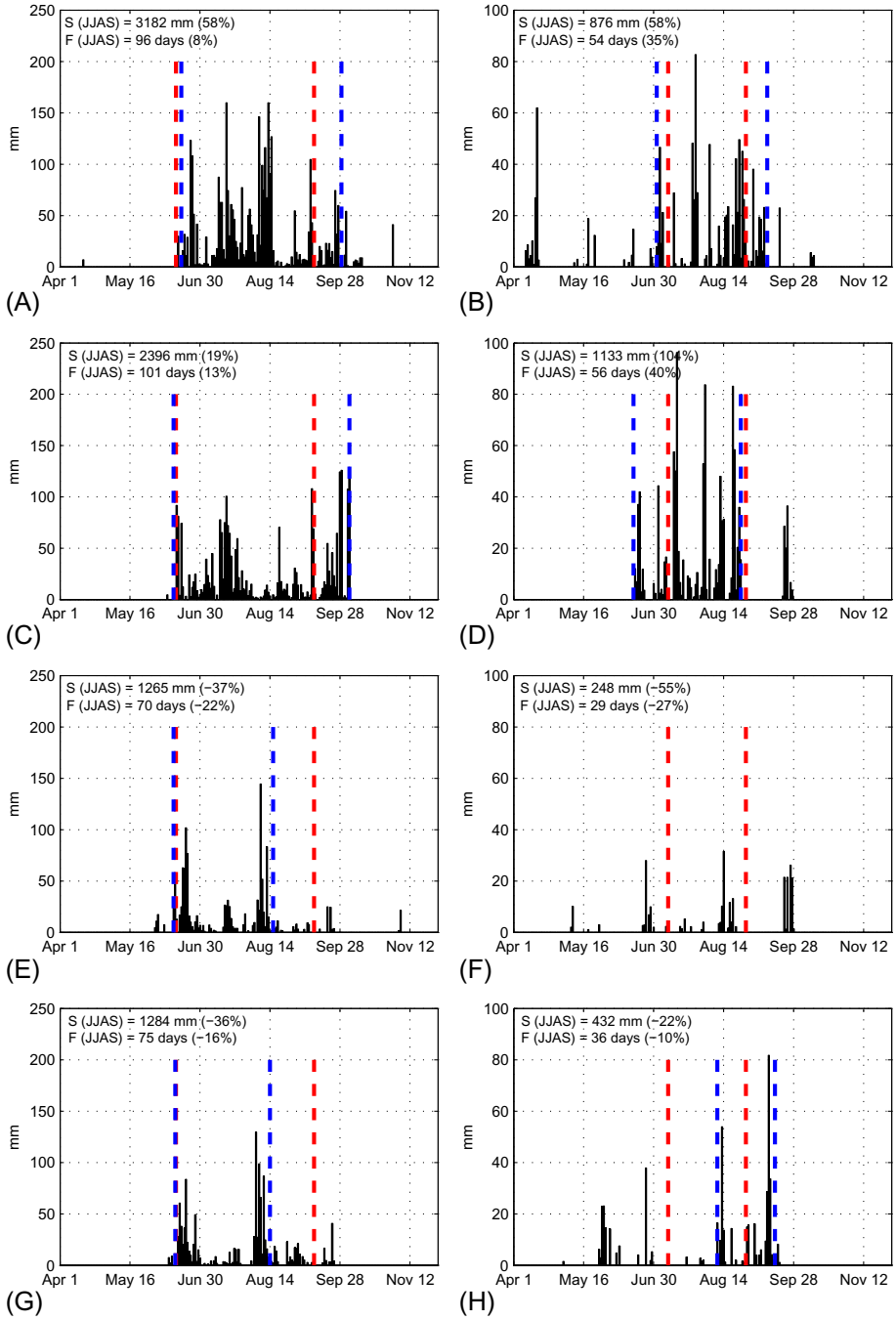

FIG. 1 Daily mean rainfall (*bars*) in millimeters for two 0.25 degrees grid points from the IMD dataset corresponding to the locations of Mumbai (left column) and New Delhi (right column) for two "good" (anomalously wet) (i.e., 1983 and 1988) JJAS monsoons and "bad" (anomalously dry) JJAS monsoons (i.e., 1986 and 2002) at the Indian scale (Sontakke et al., 2008). The *vertical dashed red lines* denote the climatological onset and end dates of the season (defined as the first and last wet day of a 5-day wet spell receiving at least the climatological amount received during 5-day wet spells (from April 1 to November 30) without any 10-day dry spell receiving less than 5 mm in the following 30 days [and previous for end days]) (Moron et al., 2017). The onset is computed from Apr. 1 and the end is computed retrospectively from Nov. 30. The *vertical blue dashed lines* denote the onset and end dates for the given seasons and locations. (A) Mumbai 1983, (B) New Delhi 1983, (C) Mumbai 1988, (D) New Delhi 1988, (E) Mumbai 1986, (F) New Delhi 1986, (G) Mumbai 2002, (H) New Delhi 2002.

to great uncertainty. A flexible alternative is to consider the sub-seasonal modulation of spatial coherence within a sliding window across the season. The window should be short enough to properly distinguish between the various stages of the monsoon, such as onset, core, and end, but long enough to filter out the shortest time scales associated with synoptic systems. It should also be short enough to capture different phases of the MJO and related phenomena, such as the northward-propagating intraseasonal oscillation (ISO) (Krishnamurthy and Shukla, 2000, 2008; Moron et al., 2012).

3.2 Sub-seasonal Modulation of Spatial Coherence Across India

The sub-seasonal evolution of spatial coherence over India is shown in Fig. 2A for $\overline{R}$, N_R over Monsoonal India (defined by black contour in panels B–E; Moron et al., 2017). Both curves have a similar evolution, with a small, short-lived peak around the mean onset date and a larger and longer one around the mean withdrawal ones (Fig. 2A). The spatial coherence of both rainfall amounts and frequency is at minimum during the core of the rainy season, when the climatological mean amounts reach their largest annual values. Such behavior has been observed using other metrics of spatial coherence, including the number of degrees of freedom (Moron et al., 2017). The spatial coherence is larger for frequency than for amounts, especially during the core of the season (thin vs thick red curves; Fig. 2A).

The four lower subpanels of Fig. 2 show the spatial patterns of the leading empirical orthogonal function (EOF) of the interannual ranks of amounts at all India grid points for four 15-day subperiods around the onset, during the core and then around the withdrawal (vertical red lines in Fig. 2A). The leading EOF pattern is dominated by the interannual variability over the core monsoonal zone, including the Western Ghats, the northern part of the Peninsula, most of the Indo-Gangetic plain, and the desert areas in the northwest. However, positive loadings are generally considerably higher around the onset (Fig. 2B) and withdrawal (Fig. 2E) than during the core of the monsoon in July (Fig. 2C) and August (Fig. 2D). The explained variance drops significantly during July-August. The interannual variations of the 15-day amount and frequency around the onset date will partly convey the anomalous advance or the delay of onset, consistent with the sharp peak in spatial coherence seen near the beginning of June in Fig. 2A. However, the more gradual increase from late August to early October may be accounted for by changes in withdrawal date, which takes more time to achieve across India than onset (Moron and Robertson, 2014) and/or the buildup of spatial coherence during the monsoon season itself. Overall, this sub-seasonal modulation suggests that a significant portion of precipitation in July-August is due to intense rainfall (Stephenson et al., 1999) embedded in small to mesoscale features, which have a smaller spatial scale than the large-scale atmospheric circulation patterns responsible for the monsoon onset and withdrawal (Moron et al., 2017).

3.3 Sub-seasonal Modulation of Spatial Coherence Over the Whole Tropical Zone

Do these findings for India generalize to the whole tropical zone (30 degrees N to 30 degrees S)? Fig. 3 shows the results of a similar analysis of the global tropics (including oceans) using

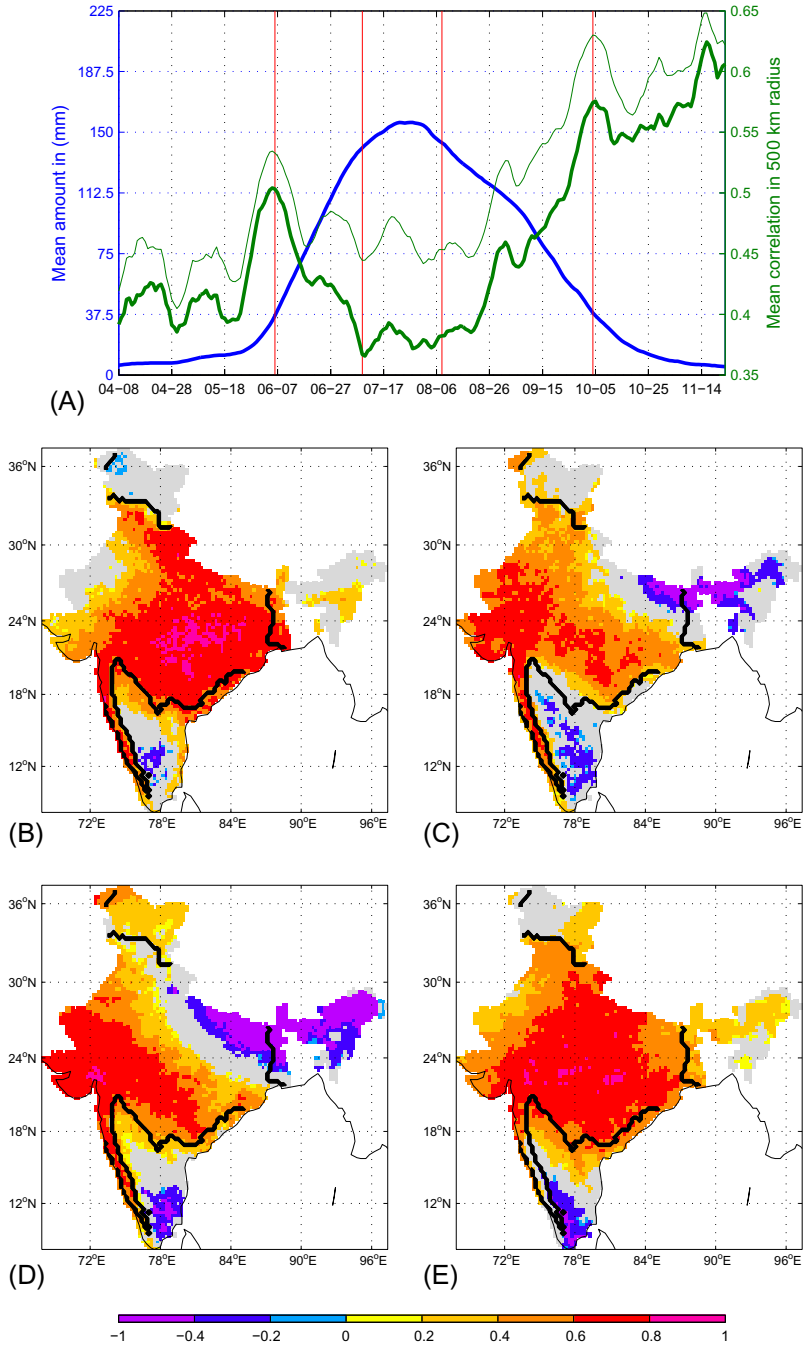

FIG. 2 See legend on opposite page.

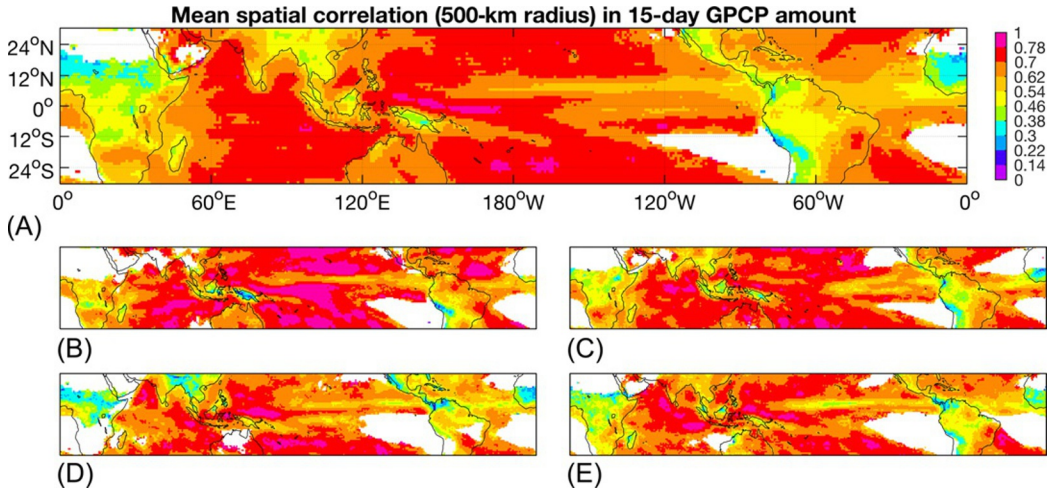

FIG. 3 Mean spatial autocorrelations in the 500-km radius for GPCP. The spatial autocorrelations are computed using ranks of amounts of rainfall between the central grid points and all grid points in a 500-km radius (including latitudes north and south of the latitudinal limits of the map). The running 15-day windows where the climatological mean amount ≪10 mm are not considered in the time average. *Blank areas* never reach the threshold of a climatological mean ≥10 mm in any of the running 15-day windows. All computations are done over Oct. 1996 to Sep. 2016. For the seasonal maps (four lower panels), all running 15-day belonging to a given season are considered (i.e., for December-February, the seventy-six 15-day periods from Dec. 1-Dec. 15 to Feb. 14-Feb. 28 are considered). (A) Year, (B) DJF, (C) MAM, (D) JJA, (E) SON.

daily rainfall amounts from GPCP. The analysis is applied over the whole calendar year using 365 running 15-day windows (top panel) and those belonging to the four usual meteorological seasons (lower panels).

The spatial coherence of 15-day GPCP rainfall anomalies is much lower over land than over the oceans in general. Area averages are given in Table 1, including for coastal areas (defined as sea ≤500 km to land) and open ocean (≫500 km from land). Some of the smallest values of spatial coherence are associated with high mean daily intensity (not shown), especially over the continents (South and Southeast Asia, Western Amazonia, etc.), although over the oceans, the equatorial eastern Indian Ocean and West Pacific warm pool ocean regions exhibit both

FIG. 2, CONT'D (A) Mean seasonal variations of spatially averaged amounts (*blue line*, left ordinate) and spatial autocorrelations of interannual ranks of mean amount (*bold red*, right ordinate) and frequency of wet days receiving ≥1 mm (*thin red*, right ordinate) in a 500-km radius in running 15-day windows (centered on the dates shown on the *abscissa*) averaged over the summer monsoonal regime (Gadgil, 2003) defined in Moron et al. (2017) from a clustering of the mean annual cycle across India and *underlined* by a black contour in panels (B)–(E). The grid points and 15-day windows receiving ≪ 10 mm in mean (over the 1901–2014 period) are excluded from the spatial average of autocorrelations. The *vertical red line* in panel (A) shows the center of the four 15-day windows used to extract the leading EOF of amount ranks in panels (B)–(E) computed over all India grid points. The EOFs are shown as loadings; that is, the correlations between the amount ranks and the leading principal component and the explained variance is noted in the panel (B) to (E) titles. Correlations that are not significant at the two-sided 95% level according to a random-phase test are indicated by *gray areas*. (A) Mean amount and spatial autocorrelation in 500-km radius [amount, frequency]. (B) EOF 1, May 30–June 13, $V = 27\%$. (C) EOF 1, Jul. 2–16, $V = 19\%$. (D) EOF 1, Aug. 1–15, $V = 23\%$. (E) EOF 1, Sep. 28–Oct. 12, $V = 29\%$.

TABLE 1

	Land	Coast	Open Sea	OLR	CMAP
Year	0.53	0.64	0.68	0.61	0.65
Dec.-Feb.	0.59	0.66	0.71	0.66	0.69
Mar.-May	0.55	0.64	0.69	0.67	0.68
Jun.-Aug.	0.49	0.62	0.66	0.70	0.67
Sep.-Oct.	0.53	0.64	0.67	0.64	0.64

Notes: *Columns 2–4: Spatial averages of the mean autocorrelation in a 500-km radius for land, coast ($\leq$500 km from land) and open sea ($\gg$500 km from land). The areas where mean rainfall over sliding 15-day windows is always $\ll$10 mm during the periods shown in Column 1 are not considered in the spatial averages; columns 5–6: pattern correlation between GPCP maps (linearly interpolated over CMAP and OLR grids) and the mean autocorrelation in a 500-km radius for the whole tropical zone. The areas where mean rainfall over sliding 15-day window is always $\ll$10 mm during the periods shown in column 1 are not considered in the pattern correlations.*

high mean intensity and strong spatial coherence (Fig. 3). Relative minima in spatial coherence coincide with the ITCZ over the northern equatorial central and eastern Pacific and Atlantic oceans (Fig. 3), where the pattern of the spatial autocorrelation tends to be primarily zonal rather than isotropic. Seasonal values of spatial coherence peak in austral summer and are at minimum during boreal summer, especially across land in the Northern Hemisphere. The seasonal modulation may be partly related to the amplitude of ENSO events peaking near the end of the calendar year. Several pockets of large spatial coherence coincide with regions of high seasonal rainfall predictability, including the Maritime Continent south of the equator around the austral summer monsoon onset in SON (September-November) (Haylock and McBride, 2001; Moron et al., 2009b), and near the boreal summer monsoon onset over the South China Sea and Philippines in MAM (March-May) (Moron et al., 2009a). Low values of spatial coherence may be partly due to cancelation between positive and negative correlations within a 500-km radius, such as those between northeast India and the Indo-Gangetic plain during the core of the rainy season (Fig. 2C and D), or they may reflect increased ascent in a deep convection center and increased subsidence nearby. Table 1 also includes the pattern correlations between GPCP spatial autocorrelations and those computed using CMAP and OLR datasets, demonstrating that the results are quite robust to the choice of rainfall or deep convection dataset. The lower spatial coherence across land versus ocean may be, at least partly, explained by the interactions promoting small-scale deep convection and involving the diurnal cycle, land-sea, and mountain-valley breezes, as well as gravity wave effects (Yang and Slingo, 2001; Slingo et al., 2003), which are far stronger over the landmasses and the coastal areas than over the open seas.

The sub-seasonal evolution of spatial coherence across the tropics is depicted in Fig. 4, and constructed by identifying the timing of the minimum and maximum of local spatial coherence conditioned on the local seasonal cycle of the 15-day mean rainfall amount. The spatial coherence and mean rainfall are low-pass-filtered using a recursive digital filter with a cut-off at 1/90 cycle/day. Most of the tropical zone is associated with a unimodal regime and two distinct rainy seasons are rare (not shown). Fig. 4 shows the timings of minimum and maximum spatial coherence during the year, considering only periods when the local climatological mean amount reaches at least 10 mm per 15 days, for all tropical grid points (panels A and D), as well

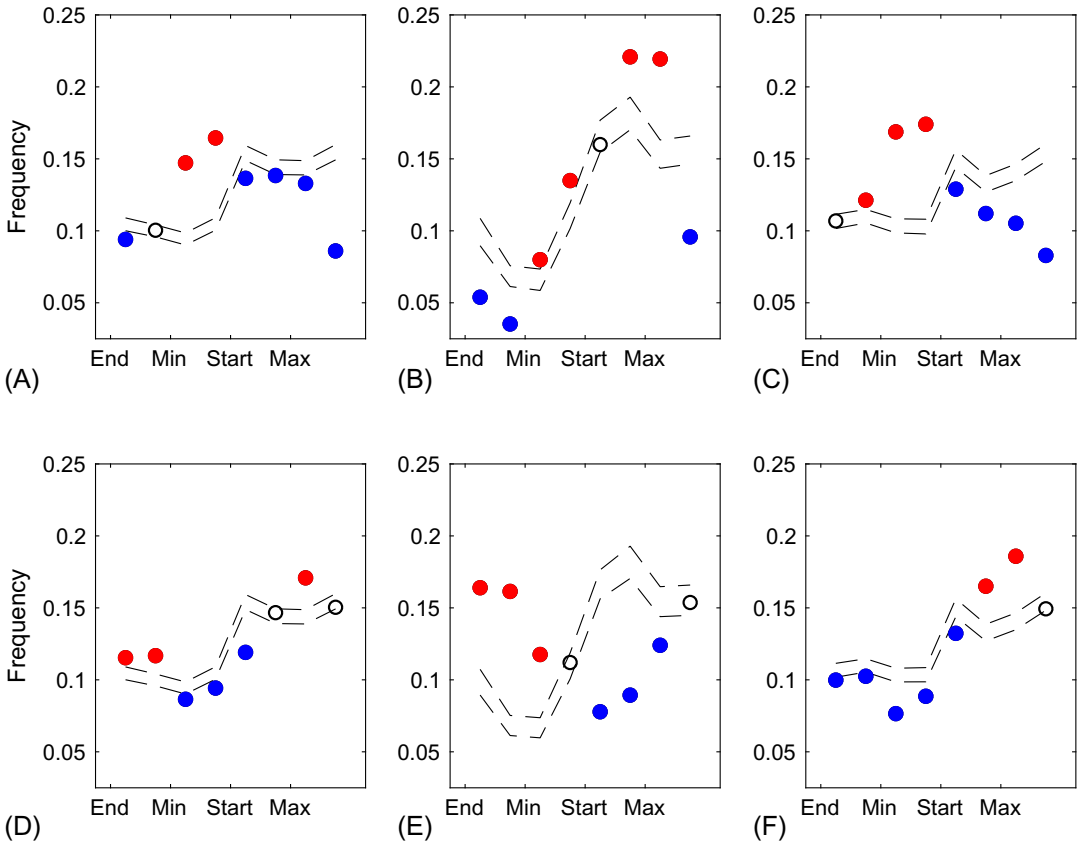

FIG. 4 Frequency of (A–C) minimum and (D–F) maximum spatial coherence (as estimated by the mean correlation between the interannual ranks of amounts on sliding 15-day windows at a central grid point and those at the surrounding ones in a 500-km radius) versus the eight phases of the mean amount of rainfall. The search is limited to the time when the mean amount $\geq$10 mm per 15-day windows, but the eight phases are computed on the whole year from the low-pass filtered (cut-off = 1/90 cycle/day) annual cycle of mean amount. The *abscissa* indicates the approximative end, minimum, start, and maximum of the local rainy season. In cases of two (or indistinct) rainy seasons, phases 2–3 correspond to the lowest annual rainfall and phases 6–7 correspond to the highest ones. The observed frequencies are indicated by *circles*, while the *dashed lines* are 95% confidence interval (CI; *dashed lines*) computed from 500 random resamplings of the time series of spatial coherence. *Red and blue circles* indicate the significant positive and negative anomalies at the two-sided 95% levels, respectively. The *first column* is for the whole tropical zone. The *second and third columns* are, respectively, for land and ocean. (A) Min. sp. coh. (all), (B) Min. sp. coh. (land), (C) Min. sp. coh. (sea), (D) Max. sp. coh. (all), (E) Max. sp. coh. (land), (F) Max. sp. coh. (sea).

as for land (panels B and E) and ocean (panels C and F) grid points separately. The minimum spatial coherence tends to occur around the time of the highest rainfall, especially over land. Maximum spatial coherence tends to occur around the onset and end of the wet season, while it is less common around the peak of the local seasonal cycle of rainfall, especially over land (Fig. 4). The sub-seasonal modulation is less clear for the ocean.

Fig. 5 shows the mean seasonal cycle of low-pass-filtered mean rainfall amount and spatial coherence for 12 regions chosen across the tropics, constructed from GPCP data. India (Fig. 5B) shows a similar behavior as those shown on Fig. 2A despite a different resolution and period covered. It is also similar to the Sahel (Fig. 5A), with the minimum in spatial coherence coinciding with the highest mean rainfall amount around early August, with the spatial coherence peaking on either side, near the start and end of the rainy season. Similar behavior is seen for most of the regions that include only land points (Fig. 5A–D, F, J, K), as well as over the North Atlantic ITCZ (Fig. 5L). Equatorial East Africa displays a different behavior during the "long rains" in MAM, with spatial coherence peaking near the start in March and then decreasing to mid-June (Camberlin et al., 2009; Moron et al., 2013, 2015a); spatial coherence is then broadly positively correlated with the seasonal evolution of rainfall during the short rains in OND (October-December) (Fig. 5E). Similar behavior is also seen for South Africa (Fig. 5K) and Amazonia (Fig. 5J), but with a weak amplitude. The remaining oceanic boxes over Western, Central, and Eastern equatorial Pacific (Fig. 5G–I) do not show a clear seasonal modulation of the (usually strong) spatial coherence.

Figs. 4 and 5 show that the spatial coherence in rainfall amount is usually low around the core of the wet season and tends to peak near its end or during its transition to dry season. They show also that the modulation is different between land and oceanic grid points, with the clearest season modulation found over land, where the monsoons are themselves strongest (note that completely dry periods (i.e., any 15-day window receiving $\ll 10$ mm in mean) are excluded from the calculations). These figures suggest that the anomalous advance and delay of the *end* of wet seasons (and secondarily their onsets) exhibit a larger scale than the core of the season, when small-scale intense, wet events may decrease the spatial coherence, at least over several continental areas such as those shown in Fig. 5A–D and F. This behavior does not appear across most of the oceans, and several continents do not exhibit the common modulation revealed by the continental regions quoted previously. As before, the explanations may involve either the variable amplitude of the impact of CCEWs (including MJO) or the variable influence of diurnal cycle, small-scale land-sea and mountain breezes and related gravity waves on local-scale rainfall. A larger (weaker) spatial coherence is expected when the first (second) processes dominate the overall variance of daily rainfall.

3.4 Skill and Spatial Coherence of S2S Reforecasts

This section presents maps of S2S model precipitation spatial coherence (Fig. 6A) and forecast skill (Figs. 6B and 7) estimated from ECMWF reforecasts, at a forecast lead time of 15–28 days (labeled "week 3 + 4"). Skill maps are constructed of the temporal correlation of anomalies (CORA), using the ensemble means of fortnight week 3 + 4 averages of the ECMWF reforecasts. The anomalies are calculated by removing the observational long-term mean of each season, as well as the models' week 3 + 4 reforecast climatology, to exclude the mean bias and model drift. ECMWF reforecasts from all the semiweekly start dates that fall in each season are lumped together as a time series in order to compare with their observed counterparts.

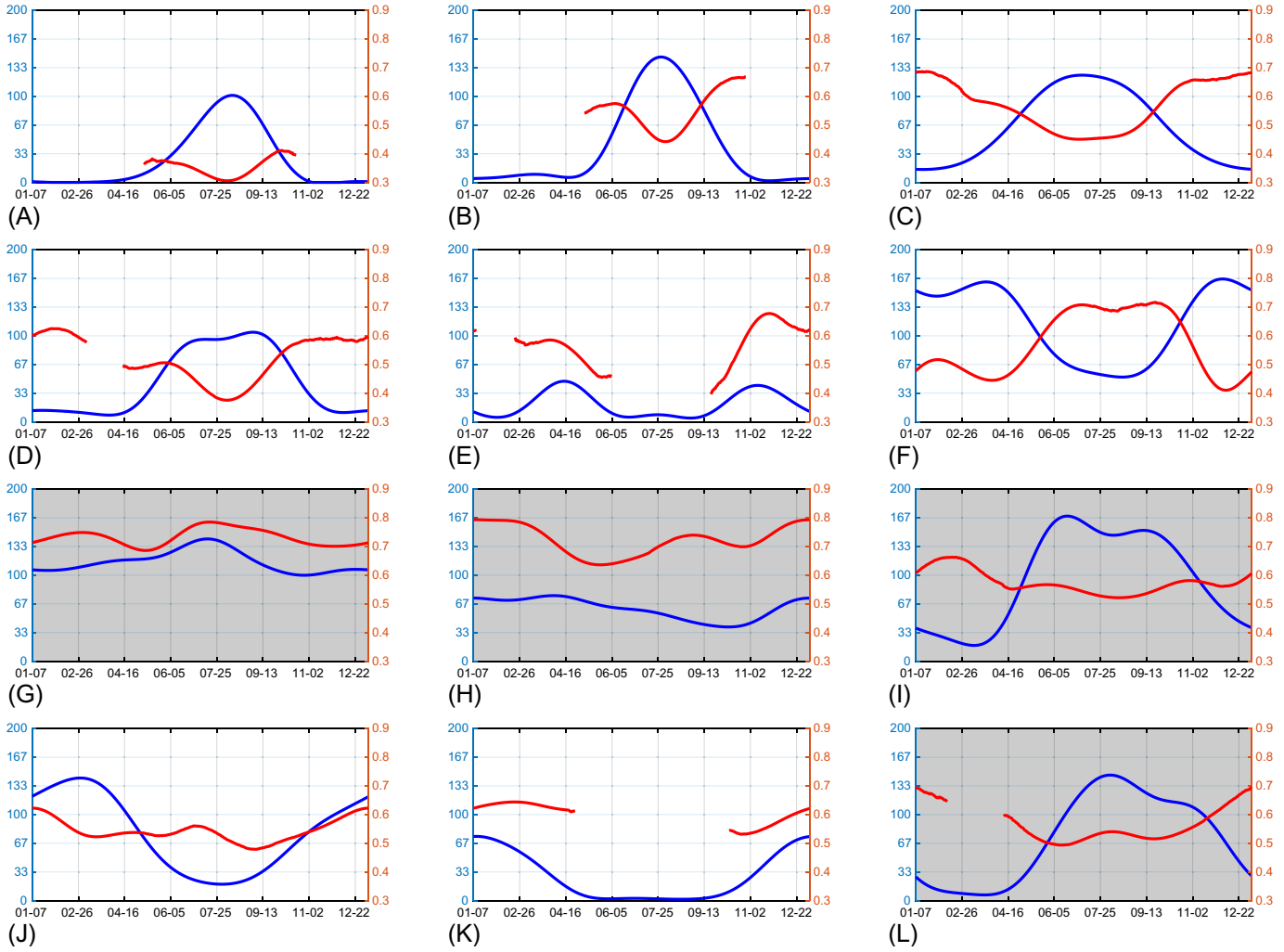

FIG. 5 Mean amount (in millimeters, *blue lines*, left ordinate) and spatial correlation of interannual ranks of amount in 500-km radius (*red lines*, right ordinate) spatially averaged over 12 tropical regions defined below. The amounts and correlations are computed using GPCP data on running 15-day windows (the center of each 15-day window is indicated on the *abscissa*) and the time series are low-pass-filtered (cut-off = 1/90 cycle/day). Data where the mean amount ≪10 mm in running 15-day windows are not used in the spatial averages. The *gray panels* use only oceanic grid points, while the *blank panels* use only landmass grid points. (A) Sahel: 12–16 degrees N, 15 degrees W-10 degrees E; (B) India: 17–30 degrees N, 72–87 degrees E; (C) SE Asia: 12–30 degrees N, 100–122 degrees E; (D) Central Am.: 12–25 degrees N, 240–270 degrees E; (E) Equat. E. Afr.: 5 degrees S-5 degrees N, 35–50 degrees E; (F) S. Indonesia: 10 degrees S-0, 90–120 degrees E; (G) ITCZ W. Pacific: 0–10 degrees N, 135–160 degrees E; (H) C. Pacific: 5 degrees S-5 degrees N, 160–210 degrees E; (I) ITCZ E. Pacific: 5–13 degrees N, 250–280 degrees E; (J) Amazonia: 15 degrees S-0, 290–320 degrees E; (K) S. Africa: 28–15 degrees S, 20–40 degrees E; (L) ITCZ Atlantic: 5–13 degrees N, 320–345 degrees E.

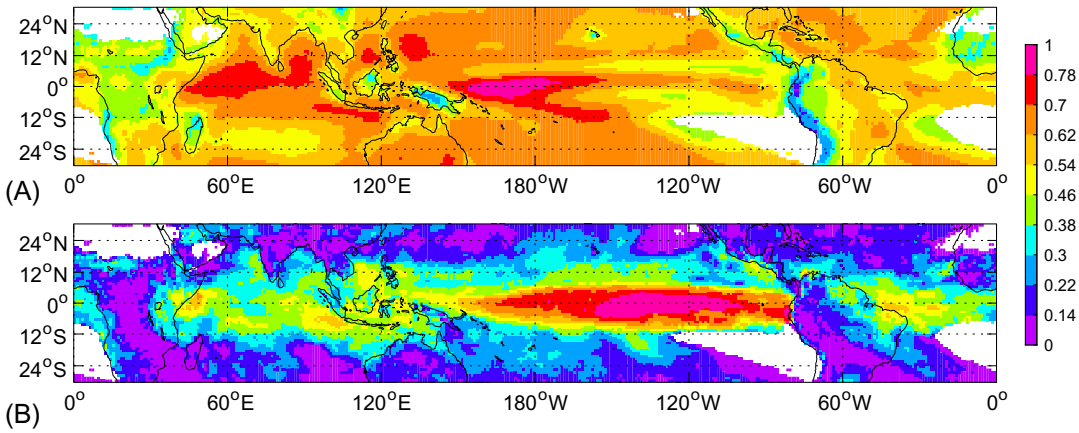

FIG. 6 (A) Mean spatial autocorrelations in 500-km radius for ECWMF week 3 + 4 reforecast precipitation anomalies. The spatial autocorrelations are computed using ranks of interannual anomalies of rainfall between the central grid points and all grid points in a 500-km radius (including latitudes north and south of the latitudinal limits of the map); (B) ECMWF week 3 + 4 anomaly correlation coefficient (CORA) for total precipitation for all year. Any CORA value >0.3 is statistically significant at the 99% CI by a one-tailed *t*-test (as only positive ACC is considered skillful). The areas receiving ≪ 10 mm in Fig. 3A are marked as blank areas.

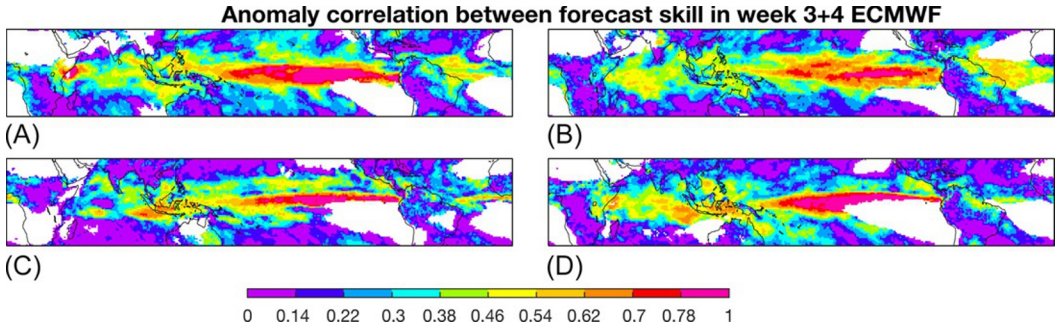

FIG. 7 ECMWF week 3 + 4 anomaly correlation coefficient (CORA) for total precipitation for (A) DJF, (B) MAM, (C) JJA, and (D) SON. Any CORA value >0.3 is statistically significant at the 99% CI by a one-tailed *t*-test (as only positive ACC is considered skillful). The areas receiving ≪ 10 mm in Fig. 3B–E are marked as blank areas.

The spatial coherence of week 3 + 4 precipitation anomalies from the ECMWF control run (Fig. 6A) (computed from the 20 reforecast years corresponding to each Monday real-time forecast start date) is very similar to the observed estimate (Fig. 3A). The pattern correlation between both maps is 0.73, reproducing the contrast between land and sea. The spatial averages of ECMWF spatial coherence equal, respectively, 0.49, 0.57, and 0.62 for land, coastal, and open ocean (close to the GPCP estimates in Table 1).

The geographical distribution of ECMWF week 3 + 4 forecast skill broadly resembles that of the spatial coherence, although the pattern correlation between them is low (0.27). The similarity in patterns is mostly related to the much higher skill over ocean than over land, especially

over the tropical Pacific due to the high local persistence there of ENSO-related SST anomalies (Li and Robertson, 2015). The geographical extent of this highly predictable belt region changes with the season (Fig. 7), being relatively narrower in JJA (June-August) (red regions in the Pacific sector of Fig. 7). Good skill can also be found over the western Indian Ocean, which extends westward to Africa, in boreal winter and fall (Fig. 7A and D). The western tropical Atlantic and the adjacent continents such as eastern Amazonia also show good skill in boreal winter and spring (Fig. 7A and B). In contrast, the eastern tropical Indian Ocean and Maritime Continent are more predictable in austral winter and spring (Fig. 7C and D).

These seasonal changes in skill are generally consistent with those in the spatial coherence as shown in Fig. 3, with higher spatial coherence being associated with higher forecast skill. This is especially clear over the Maritime Continent, where low skill over the islands tends to coincide with lower values of spatial coherence there compared to the surrounding ocean. However, this correspondence between forecast skill and spatial coherence of rainfall anomalies is less clear elsewhere. For example, skill is often high over the central and eastern Pacific ITCZ and around the Atlantic ITCZ despite relatively low spatial coherence. The lack of skill over most continents for week 3 + 4 forecasts indicates that much room remains for progress in dynamical models in capturing sub-seasonal variability in the tropical rainfall. However, the comparison with spatial coherence suggests that some of these regions are often intrinsically less predictable.

4 DISCUSSION AND CONCLUDING REMARKS

Most previous predictability studies of tropical rainfall have focused primarily on interannual variations of seasonal rainfall amount, typically associated with the atmospheric response to slow boundary forcings, primarily ENSO and other coupled ocean-atmosphere modes of variation. The temporal summation across the season smooths the characteristics of individual rainfall events, thus emphasizing the impact of systematic (i.e., near-constant in time) "slow" forcings. The progressive temporal summation across the season also increases the spatial coherence of rainfall anomalies due to the near-constant modulation of either intensity, size, or frequency of the instantaneous wet events across a region. We focused here on quantifying the spatial coherence of sub-seasonal (2-week average) rainfall anomalies, allowing different sub-seasonal stages during a wet season to be distinguished. The spatial coherence is estimated with the mean spatial autocorrelation in a 500-km radius around each grid points. These observational sub-seasonal estimates of predictable spatial scale are then compared with estimates of forecast skill from ECMWF week 3+4 reforecasts from the S2S database.

The analysis of 0.25 degrees daily Indian rainfall illustrates the fact that two "good" monsoons (1983 and 1988) and two "bad" monsoons (1986 and 2002) at the Indian scale exhibit very different sub-seasonal evolutions at the local scale (Fig. 1). The spatial coherence of biweekly rainfall anomalies peaks near onset and withdrawal of the summer monsoon and reach its minimum during the core of the rainy season, in JA (Fig. 2A). This illustrates an example where the seasonal (JJAS) amount—which is dominated by the largest rainfall in JA—may blur smaller predictable signals in June or September. The leading EOF of ranks of

TABLE 2

	All	Land	Coast	Open Sea
Year	0.30	0.20	0.31	0.34
Dec.-Feb.	0.33	0.23	0.32	0.37
Mar.-May	0.31	0.21	0.31	0.36
Jun.-Aug.	0.27	0.19	0.30	0.29
Sep.-Oct.	0.29	0.21	0.32	0.32

Notes: *Columns 2–5: Spatial averages of the mean skill for weeks 3 + 4 for all tropics, land, coast (<500 km from land) and open sea (>500 km from land). The areas where mean rainfall over sliding 15-day windows is always ≪ 10 mm during the periods shown in column 1 are not used in the spatial averages.*

amounts in JA (Fig. 2C and D) shows an out-of-phase pattern between most of the peninsula and western India on one hand and the Himalayan foothills and northeastern parts of India on the other hand, which may be viewed as the fingerprint of the main intraseasonal mode of variation (i.e., ISO; Krishnamurthy and Shukla, 2000, 2008; Moron et al., 2012), but the largest loadings ≥ 0.6 also cover a smaller area in JA than around the onset (Fig. 2B) and withdrawal (Fig. 2E) stages of the boreal monsoon. This sub-seasonal modulation with the smallest spatial coherence recorded around the core of the wet season appears to be rather general across the tropical continents (Fig. 4B and E), even if there are some exceptions (such as Equatorial Eastern Africa, Southern Africa, and Amazonia; Fig. 5E, J, and K). Moreover, the spatial scales are systematically larger across the ocean than over the landmasses (Figs. 3, 4, 6A and Tables 1 and 2). This land-sea contrast is also seen in S2S forecast skill (Fig. 6B and Table 2).

Regarding the general spatial difference between land and ocean, we can make a first hypothesis: area of subdaily wet events defined as contiguous wet grid points or those recording deep convection above a given threshold are systematically smaller over continents than over oceans (Ricciardulli and Sardeshmukh, 2002; Smith et al., 2005; Dai et al., 2009; Trenberth et al., 2017). It is also well established that the impact of the diurnal cycle and related processes, including land-sea and mountain-valley breezes and gravity waves (Yang and Slingo, 2001; Slingo et al., 2003), is stronger across continents and leads to shorter/smaller wet events across continents than over oceans. An ocean imposes, by definition, an homogeneous large-scale boundary forcing on the atmosphere, and the diurnal cycle is reduced due to thermal inertia. An open question is the interaction between the strong diurnal cycle and the modulation provided by various CCEWs (Wheeler and Kiladis, 1999; Lubis and Jacobi, 2015). For example, the amplitude of CCEWs decreases rather radically from the Gulf of Guinea to the Central Africa in boreal spring (Kamsu-Tamo et al., 2014). In that context, a strong impact of larger wet patterns, as tropical depressions or tropical-temperate-trough (TTT) systems (as those related to the South India Convergence Zone in southeast Africa and to the South Atlantic Convergence Zone in eastern Brazil), or where the CCEWs strongly affect the intensity and occurrence of local-scale rainfall may lead to a larger spatial scale of 15-day rainfall anomaly. A discrepancy is then possible between spatial coherence and predictability since a larger daily or subdaily wet pattern is not necessarily predictable at the S2S timescale. For example, a rather large spatial coherence over subtropical southern Africa (Figs. 3A, 6A), especially in

DJF (December-February) (Fig. 3B) does not match with an high predictability (Fig. 7A) during this season. In that case, the larger spatial coherence may be primarily an effect of TTTs generating a large wet pattern at the daily time scale (Macron et al., 2014) but it may be not really predictable by ECMWF with a lead time of 2 weeks. On the contrary, the combination of a large spatial coherence and high S2S predictability in ECMWF (Figs. 3 and 6), as the one over equatorial East Africa in SON (Figs. 3E, 5E, and 6) may primarily reflect a significant impact of MJO synchronizing efficiently the local-scale rainfall and being predictable at the S2S timescale (Pohl and Camberlin, 2006; Berhane and Zaitchik, 2014). The combination of relatively large spatial coherence with high S2S predictability is also observed over northeastern South America, and smaller continental pockets such as northern Australia or southern China.

Regarding the temporal modulation of spatial coherence (and S2S forecasts), several hypotheses may be emphasized. First, the seasonal modulation of the large-scale source of predictability could be a trivial source. We have shown here that the spatial coherence peaks and has high predictability in DJF, in phase with the usual peaks of ENSO events (Rasmusson and Carpenter, 1982; Ropelewski and Halpert, 1987, 1996). But the minimum spatial coherence is not observed in MAM at the usual time of shift between warm and cold ENSO events, but rather in JJA, when the ITCZ reaches its farthest location from the equator and when it is located over several landmasses in Central America, and between Western Sahel-Sudan and Southeast Asia. The exact role of the latitudinal distance of the main heating source from the equator and of the area of these heating sources on the scale and predictability of atmospheric anomalies remains to be established. A second source may be the role of extreme wet events that are known as significantly decreasing the spatial coherence and predictability of seasonal amounts in India (Stephenson et al., 1999; Moron et al., 2017). A third source of subseasonal modulation could be related to the multiscale interaction involving the slow forcings, the basic atmospheric state, and the diurnal cycle across the seasonal cycle. For example, it has been shown that warm ENSO promotes regional-scale subsidence over the Maritime Continent, but the interaction between this signal and local-scale rainfall varies over time. During warm ENSO events, the regional-scale, anomalous, low-level easterlies tend to counteract the usual monsoon flow and thus promote the occurrence of a quiet weather type (Moron et al., 2015b) characterized by anomalously weak, low-level winds during the core of the wet season (i.e., DJF). These weak winds emphasize the role of the diurnal cycle that lead to localized positive rainfall anomalies restricted to small parts of the islands despite the regional-scale subsidence anomalies (Qian et al., 2010). In that case, the spatial fragmentation of the rainfall anomaly field occurs during the core of the wet season, when the westerly basic flow is strong enough to be almost canceled by the anomalous easterlies associated with the ENSO forcing. Such varying interaction may be possible over other archipelagos as the Caribbean basin.

Finally, the fact that the spatial coherence usually peaks either near onset or withdrawal needs further analysis. First, the anomalous delay or advance of regional-scale onset or withdrawal may be the primary signal of the peak in spatial coherence around these times. We can also hypothesize a role played by land-atmosphere interaction: For most of the tropical continents except for some equatorial areas with very long, or even constant, moist conditions, the soils are fully dried up during 6–10 months. We can then hypothesize that the onset of the season over the continents is mostly forced by large-scale atmospheric circulation because soils cannot provide any local moisture that can trigger deep convection. In other words, a

tropical landmass at the end of the dry season would wait for the favorable conditions associated with a combination of any source of slow and S2S predictability and large-scale phenomena conveying enough moisture to get the first rains. The intense warming of the dry surfaces lowers the static stability of the lower atmosphere. It is possible that the combination of local dry and moist instability conveyed by large-scale atmospheric circulation is very efficient close to the start of the season. The first rains in India are indeed very intense (Moron et al., 2017). The first wet events either related to local-scale thunderstorms or MCS would have two effects: (i) moistening the soils and starting the local recycling of water, which is indeed a positive feedback (Meehl, 1997; Douville et al., 2001, 2007); and (ii) increasing the heterogeneity of surface temperatures and humidity at the regional scale. As time goes by during the rainy season, the second effect will progressively vanish due to the different locations (or tracks) of wet events (from thunderstorms to MCSs) due to atmospheric circulation (including S2S phenomena). It could provide a partial explanation of the increase of spatial coherence toward the end of the season, at least over India (Douville et al., 2001, 2007), and which seems to be rather general across most of the tropical continents. This effect may be superimposed on the increasing power of ENSO events, at least for the northern tropics.

CHAPTER

4

Identifying Wave Processes Associated With Predictability Across Time Scales: An Empirical Normal Mode Approach

Gilbert Brunet[*], *John Methven*[†]

[*]Meteorological Research Division, Environment and Climate Change Canada, Dorval, QC, Canada [†]Department of Meteorology, University of Reading, Reading, United Kingdom

OUTLINE

https://doi.org/10.1016/B978-0-12-811714-9.00004-8

1 INTRODUCTION

It can be demonstrated that predictive skill ranges from weeks out to seasons when the forecast metric involves the statistical treatment of weather variables. Typically, the range of regional forecast skill increases with the scale of the region considered, as well as the length of the time window used for verification (see Chapter 2). Comparisons can be made in terms of averages over forecast lead times, over ensemble members, or, in a probabilistic sense, using ensemble forecasts. These extended-range forecasts go beyond the limit of predictability for point forecasts (the value of an atmospheric variable at a particular time and location). Such forecasts depend on the nature of the variable being forecast and the phenomena dominating its fluctuations. For example, geopotential on a pressure surface is a smooth field dominated by synoptic-scale (or larger) weather systems and is predictable out to 7–15 days, depending upon the flow configuration. Finer-scale fields, like vorticity, have shorter predictability limits. For example, Frame et al. (2015) showed that the predictive skill for the strike probability of cyclonic vorticity centers, within a given radius of locations in the Euro-Atlantic sector, increases with feature intensity and scale. Precipitation typically has a much shorter limit to its predictability, owing to its dependence on vertical motion and convective-scale features in the flow.

There are many plausible sources for sub-seasonal to seasonal (S2S) predictability, including slowly varying boundary conditions for tropospheric weather systems: coupling with the ocean (see Chapters 5 and 9), land surface processes (see Chapter 8), the cryosphere (see Chapter 10), and the stratosphere (see Chapter 11). However, sub-seasonal regimes are characterized by large-scale patterns of variability that exhibit internal variability over long time scales, and they can be oscillatory in nature or characterized by long-range teleconnections. Consider three contrasting examples:

1. The weather in the tropics and the extratropics is influenced by the phase of the Madden-Julian Oscillation (MJO). The MJO is the dominant tropical mode of sub-seasonal variability, which propagates eastward along the equator but has strong remote influences (see Chapters 5 and 7).
2. Unprecedented extreme summer precipitation events in western Europe have occurred in the last decade, and they have been associated with quasi-stationary Rossby wave patterns on the midlatitude jet stream (Blackburn et al., 2008), prompting new theories of resonant excitation of Rossby modes (Petoukhov et al., 2013; Coumou et al., 2014).
3. The Russian heatwave that occurred in 2010 was associated with a persistent midlatitude blocking regime (Dole et al., 2011).

In these examples, a distinct dynamical phenomenon is involved, and the properties of that phenomenon influence the weather in a predictable manner. Here, we focus on isolating oscillatory phenomena or slowly propagating modes responsible for enhanced predictability in the S2S range. In addition to decoupling the phenomena from the data, it is important to gain insight into their intrinsic dynamical properties and interactions in order to anticipate how they will contribute to predictability. Also, in the context of a changing climate, if we understand how modes of variability relate to the climate state, we will be able to anticipate how variability and predictability may vary with climate change.

Energy spectra in the wave-number space (calculated by transform of spatial distributions and averaging over many realizations) show a smooth continuum from the planetary to the kilometer scale, indicating that there is no spectral gap distinguishing large-scale phenomena from smaller-scale ones. Similarly, frequency spectra exhibit a smooth continuum from seasons to hours. However, statistical techniques that utilize both spatial and temporal information, such as the popular empirical orthogonal function (EOF) technique, show that covariance in time is typically dominated by large-scale patterns of variability. The EOF technique for a given variable maximizes the variance, which is explained by a series truncation of fixed length.

A remarkable property of an EOF mathematical construction is that EOF patterns are orthogonal to one another (i.e., the integral over the domain of two different EOF patterns multiplied together equals zero) and their corresponding time series are orthogonal as well, making them a complete basis that can be used to project variability from a discrete data set. Hence, the first EOF is the spatial pattern that explains most temporal variance; the first and second EOFs form the two-dimensional (2D) orthogonal basis that explains the most variance, and so on for higher terms in the series. The disadvantage of using a purely statistical approach is that the spatial structures obtained (the EOFs) and their corresponding principal component time series do not have distinct physical properties; therefore, it is difficult to anticipate their behavior beyond the limits of the time series examined.

A relatively unexploited approach is to combine conservation laws derived from consideration of atmospheric dynamics with the orthogonality approach to identify distinct patterns of variability that can be linked to characteristic physical properties (e.g., an intrinsic frequency or a phase speed). This approach can be called *empirical normal mode (ENM) analysis.* Just as an analysis of the shape and physical construction of a bell can be used to anticipate the frequency at which it will ring when struck, the spatial structure of an ENM can be used to predict its intrinsic frequency or phase speed from conservation properties. If it can be shown in general that a small number of such modes dominate S2S variability, as was demonstrated by Brunet (1994) for the 315 K isentropic surface, a huge reduction in the dimensional size of the system to be solved will result. Hence, the ENM basis could be the natural one to use to study waves and weather regimes in low-order dynamical systems, as discussed in Chapter 6.

Other disturbances may perturb the modes over a wide range of frequencies or stochastically, which causes the observed frequency spectrum to resemble a continuum. Therefore, knowledge of their intrinsic frequencies provides potentially useful information. The well-known fluctuation-dissipation theorem (FDT) describes how the time-average response of a dynamic system to random perturbations will resemble the structure of the slowest (longest-time-scale), unforced mode of variability. For example, Ring and Plumb (2008) studied the response of the Southern Annular Mode to forcing by drawing on the FDT. They used the principal oscillation pattern (POP) analysis developed by Hasselman (1988) and Penland (1989), which relies on the calculation of lag covariances to obtain the temporal behavior of spatial patterns. An advantage of the ENM technique is that lag covariances are not required because the frequency information stems from the dynamical properties.

A popular approach used in both weather and climate sciences is *composite sampling* to study long geophysical time series. With this method, statistical properties (e.g., average and standard deviation) of similar segments of a time series are examined, and a given segment is included in the composite if a characteristic event occurs within it. There is a vast body

of literature describing such approaches; these studies use atmospheric reanalysis time series to link weather-climate events (e.g., Molteni et al., 1988; Robertson and Metz, 1989; Cotton et al., 1989; Vautard, 1990; Ferranti et al., 1989; Lin and Brunet, 2009). For example, Asaadi et al. (2016a, 2017) used this approach to identify fundamental dynamical and physical processes related to hurricane genesis. They showed that the coexistence of an African easterly wave nonlinear critical layer and a region of a weak meridional potential vorticity (PV) gradient over several days might be a major factor determining whether tropical disturbances develop into hurricanes. This finding answered the long-standing question of why only a small fraction of African easterly waves contribute to hurricane genesis. The study showed a way how S2S variability can modulate the hurricane season.

In general, the utilization of models of varying complexity is also needed for understanding and identifying sources of predictability. For example, with regard to sub-seasonal variability, this approach made it possible to demonstrate for the first time a two-way linkage between the Madden Julian Oscillation (MJO) and the Arctic Oscillation (AO) in a simplified general circulation model (GCM), numerical weather prediction (NWP) systems, and observations. These studies have pointed the way toward improving S2S predictive skill and provided examples of how global teleconnections are influencing regional weather in more complex modeling systems (Lin et al., 2007, 2009, 2010a; Lin and Brunet, 2011); also see Chapter 7.

A key message arising from this work is that even though the atmosphere is chaotic and stirring by large-scale waves and eddies generate fine-scale structures in air masses and conserved properties such as PV, the large-scale dynamics may be closer to linear behavior than expected. Characteristic spatial patterns that vary slowly but are not periodic in nature are often treated as oscillations. Teleconnections typically fall in this category. There are many situations in which interactions between wave phenomena can be identified and explained mechanistically using linear dynamics. The purpose of this chapter is to provide a theoretical and statistical framework for studying the sub-seasonal predictability in observational and model data using these concepts in an integrated manner.

Section 2 introduces the components of the framework, including the notional partition between a background state and perturbations to it, the concept of wave activity conservation, and the implications of conservation for modes of variability. Section 3 outlines the ENM technique and presents some implications for the behavior of perturbations that can be deduced from the approach. Several applications of the approach to global atmospheric data are presented to illustrate the technique and its potential. Conclusions are presented in Section 4.

2 PARTITIONING ATMOSPHERIC BEHAVIOR USING ITS CONSERVATION PROPERTIES

There are two major aims of the approach to understanding atmospheric variability presented here:

- Development of a theoretical framework that is capable of isolating a slowly varying component of the atmosphere that is influenced by slow processes, such as radiative forcing, and described by well-known equations. This phenomenon may be considered to be linked with climate.

- Development of a technique to isolate coherent, dynamical modes of variability from observed global data and models. These modes have intrinsic properties that can be deduced from theory.

In spite of the fact that there is a continuum of complex behavior (including nonlinear interactions) across scales, our goal is to take the underpinning theory as far as possible in terms of isolating processes from observations and representing them in forecasts. The robustness of the approach will be tested with global reanalysis data. Possible applications that result from the deductions about forcing of variability, dynamical processes and wave resonance also will be discussed.

Areas where this approach could be useful include the following:

- Anticipating how variability might change with changing climate by increasing our understanding of the properties of dynamical modes of variability and their dependence on the background state
- Identifying the physical links between large-scale modes of variability and high-impact weather that typically occurs on smaller scales
- Using these links to forecast the likelihood of high-impact weather, even though representing high-impact weather itself in models may be very challenging
- Diagnostics to identify model errors in the representation of dynamics

Examples include the risk of extreme precipitation that is conditional on the phase of the MJO and persistent midlatitude weather extremes associated with a particular phase of quasi-stationary Rossby waves, such as the extremely wet summers that occurred in western Europe in 2007 and 2012 (Blackburn et al., 2008; de Leeuw et al., 2016), or the Russian heatwave in 2010 (Dole et al., 2011).

2.1 Partitioning Variability: Background State and Wave Activity

Typically, atmospheric variability is identified through statistical analysis that does not explicitly use the properties of atmospheric dynamics. A starting point is to identify an anomaly: it could be some form of readily recognized coherent structure in the atmosphere (such as a tropical cyclone), but more often it is obtained by subtraction of some form of mean state to define perturbation fields at every point in the model:

$$q' = q - q_0 \tag{1}$$

These perturbations are then analyzed statistically. In such an approach, the definition of perturbations and their properties clearly depends on the mean state, q_0, that is chosen.

The three most popular approaches for defining mean state are as follows:

1. *Global average*. In this case, the system can be described by the global integral of the evolution equations, but all the dynamics are contained entirely within the perturbations that dominate the atmospheric response to forcing (e.g., the global temperature response to greenhouse gas forcing or volcanic eruption).
2. *Eulerian time average (at fixed locations)*. The mean state can readily be calculated from data, but it is not a complete solution of the governing equations. Forcing from eddy fluxes

must be added. All the time dependence rests within the perturbations. One example of this approach is forecasting a regional seasonal temperature anomaly in contrast to numerical weather prediction (NWP) of the total temperature field.

3. *Eulerian zonal average (average around latitude circles)*. Although it is often used in dynamic meteorology, as in the wave-mean flow interaction problem, this approach has a disadvantage: The mean can vary as quickly as the perturbations because it may be changed by adiabatic eddy fluxes.

Research in the 1960s, 1970s, and 1980s showed that the evolution of the mean meridional circulation (a zonal mean, usually time-filtered) depends crucially on the variables chosen to describe the data (Andrews et al., 1987). In particular, there is a very marked difference in the mean state deduced by averaging pressure-level data (the easiest coordinate to use in assimilating observational profile data), compared with averaging along isentropic levels (surfaces of constant potential temperature). The key reason is that potential temperature is materially conserved following adiabatic motion, so isentropic-coordinate averages partition diabatic behavior from adiabatic behavior. In contrast, vertical motion across pressure surfaces can occur through both adiabatic and diabatic processes. A major example is the Ferrell cell, which appears in the midlatitudes as a thermally indirect circulation in the pressure-coordinate average, but is absent in the isentropic-coordinate average (Townsend and Johnson, 1985). The origin of this structure is predominantly adiabatic motion along sloping isentropic surfaces within midlatitude weather systems.

An extension of the isentropic-coordinate approach involves using coordinates where two of the variables are properties that are approximately materially conserved. The most common example is to use PV in conjunction with potential temperature (e.g., Nakamura, 1995). If motions are adiabatic and frictionless, then the PV and potential temperature cannot be modified along trajectories. If surfaces of constant potential temperature and PV intersect, those intersections must be transported around by the fluid as material lines. For example, the midlatitude tropopause is often described as a particular PV surface (2 PVU), and the position of the tropopause on each isentropic surface that intersects it is then stirred by the fluid motion along that surface (Hoskins et al., 1985). However, for conservative motion, mass cannot be transported across isentropic surfaces or PV surfaces, nor can it be transported across the intersections between the surfaces. This is a strong constraint on fluid behavior.

In a conceptual sense, we can describe the entire atmosphere in conserved variable coordinates (i.e., a 2D plane with potential temperature and PV as the axes), where motion of atmospheric mass is possible only through the action of diabatic or frictional processes. This framework is described as a *modified Lagrangian mean (MLM)* (McIntyre, 1980). A true Lagrangian mean would require calculation of the trajectories of all air parcels and some form of time-averaging along trajectories and over trajectories from similar initial conditions. In contrast, the MLM state is obtained by using the approximately conserved variables as markers of air masses, and therefore tracers of fluid motion. This has a distinct advantage. The time-dependent winds stir tracers so that fine-scale structures are developed through chaotic advection, and there is a cascade to ever-finer-scale structures in the absence of nonconservative processes. However, real tracers including chemical constituents, as well as PV and potential temperature, are subject to nonconservative processes that act to dissipate the smallest features (halting the scale cascade) and also act to maintain large-scale contrasts

(so that the entire atmosphere does not become well-mixed). The MLM approach takes advantage of this property.

Another crucial aspect of introducing a partition between a background state and perturbations about that state is that it is necessary to predict the evolution of both components using the equations of motion and thermodynamics. First, we will consider the background state (the evolution of the perturbations will be considered in Section 2.2). An important way to predict background state evolution is to use the integral conservation properties of the full state. The definition of the MLM state uses two material conservation properties. From this, it is possible to deduce that mass cannot cross surfaces of constant potential temperature (θ) or constant Ertel PV (q) if the flow is conservative (adiabatic and frictionless). Therefore, the mass enclosed in a volume bounded above and below by two neighboring isentropic surfaces ($\Delta\theta$) and laterally by a PV contour (with value Q) must be conserved:

$$M(Q,\theta) = \frac{1}{\Delta\theta}\iiint r\, dA\, d\theta \tag{2}$$

where r is the density in isentropic coordinates and A is the area enclosed.

Kelvin's circulation theorem also states that the circulation within any closed material contour on an isentropic surface is invariant if the flow is adiabatic and frictionless. A set of circulation integrals can be defined by the integral of the tangential absolute velocity (in an inertial frame) around all closed PV contours (i.e., varying Q):

$$C(Q,\theta) = \oint_{q=Q} \underline{u}.d\underline{l} = \langle u_t \rangle L_Q \tag{3}$$

where the average tangential speed of the flow around the circuit, $\langle u_t \rangle$, must depend inversely upon the length of the circuit, L_Q. Although the mass enclosed by the circuit is invariant, the length of its boundary depends upon the degree of contortion of the PV contour by the flow. Therefore, the average speed and local velocity must depend on the shape of PV contours. To specify the background state and perturbations to it completely, it is necessary to make an assumption about the shape of the PV contours that define the background state.

One option is to define the MLM background state as zonally symmetric and to require it to obey the same equations of motion as the full flow. This can be achieved by an adiabatic rearrangement, in which the mass and circulation enclosed by every PV contour in isentropic layers are the same as in the full state. The geometry of the contours is made to be concentric around latitude circles (McIntyre, 1980; Methven and Berrisford, 2015). Using Stokes's theorem, the circulation also can be expressed as a volume integral of PV:

$$C(Q,\theta) = \frac{1}{\Delta\theta}\iiint rq\, dA\, d\theta \tag{4}$$

because the absolute vorticity on an isentropic surface is given by rq.

Consider a thought experiment where there is a distorted polar vortex in an isentropic layer characterized by uniform high PV (value Q) inside the vortex and zero PV outside. For illustration purposes, perturbations in isentropic density along the layer are assumed to be small compared with the mean density, R. In this case, the circulation within the wavy

contour is approximately RQA, where A is the area enclosed. Therefore, the background state zonal flow around the vortex edge is

$$u_0 = \frac{RQA - C_p}{L_0} \tag{5}$$

where L_0 is the length of the latitude circle encompassing a vortex of area A that is concentric about the pole in the background state. It is usual to define an equivalent latitude for the contour such that $L_0 = 2\pi a(\pi/2 - \phi_e)$ and $A = 2\pi a^2(1 - \sin\phi_e)$, where a denotes the Earth's radius. The area integral of the vertical component of the planetary vorticity ($2\Omega \sin\phi$) within the background state contour is $C_p = \Omega.\, 2\pi a^2(1 - (\sin\phi_e)^2)$. It is subtracted to obtain the zonal flow in the rotating frame of the Earth. The zonal average of the flow around latitude ϕ_e is then given by

$$[u] = \frac{RQ(A - B) - C_p}{L_0} \tag{6}$$

where B is the area within the latitude circle that is external to the disturbed vortex where the PV is zero. Two properties are immediately apparent. If there is any disturbance, we expect $[u] < u_0$ because $0 < B < A$, and we also can anticipate that the Eulerian zonal average $[u]$ will fluctuate with the disturbance amplitude, as characterized by the area occupied by low PV ridges (B).

The flow and density can be obtained by inversion of the PV distribution (Methven and Berrisford, 2015). In this way, we can define the distribution of dynamical atmospheric variables and their evolution. Because the PV and θ distributions of the MLM state are zonally symmetric, the zonal flow obtained from PV inversion must be symmetric as well. Furthermore, because the zonal flow is parallel to the PV contours along isentropic surfaces, there can be no change associated with advection. Without additional approximations, the solution to the primitive equations on the sphere in this situation is a state that is in hydrostatic and gradient wind balance. Because the zonal integral of the full state conserves zonal angular momentum, so must the zonally symmetric background state due to translational invariance in the zonal direction. The perturbations also must obey a pseudoangular momentum conservation law, as described in Section 2.2.

An alternative would be to define the background as a strictly steady state. In this case, the time invariance implies energy conservation of the background, as well as global energy conservation of the full state, and the perturbations obey a pseudoenergy conservation law. However, there are several disadvantages to this approach. The background state is not exactly in balance if the flow is zonally asymmetric, so the PV distribution cannot, in general, be inverted to obtain the flow and density. Furthermore, except in a special situation where the flow happens to be parallel to PV contours everywhere, the background could not be steady without the continuous action of forcing introduced into the evolution equations.

Therefore, a compromise must be made. Either we identify the background state with zonal symmetry, in which case even stationary waves and zonal variations in climate must be regarded as part of the perturbation field, or we identify the background with a steady state (time symmetry), in which case the evolution of the background is not considered (by definition), and the maintenance of the background also would require a forcing term

in the equations. Most people would define the term *climate* using some notional component that is slowly varying and inherently large-scale, but this approach loses the advantages of precise symmetry in space or time.

In the analysis of perturbations that follows in this chapter, we will use a zonally symmetric background state, but also assume that it evolves much more slowly than the perturbations. As explained previously, this is approximately true for the MLM state because it can evolve only through nonconservative processes, and these modify global circulation only slowly. Fig. 1 illustrates the evolution of the MLM state for the Northern Hemisphere in June 2007, obtained using the equivalent latitude iteration by PV inversion (ELIPVI) method of Methven and Berrisford (2015).

It is immediately apparent that $u_0 > [u]$, and also that $[u]$ varies more rapidly, as predicted previously. The solid black contour on both plots marks the tropopause (PV = 2 PVU;

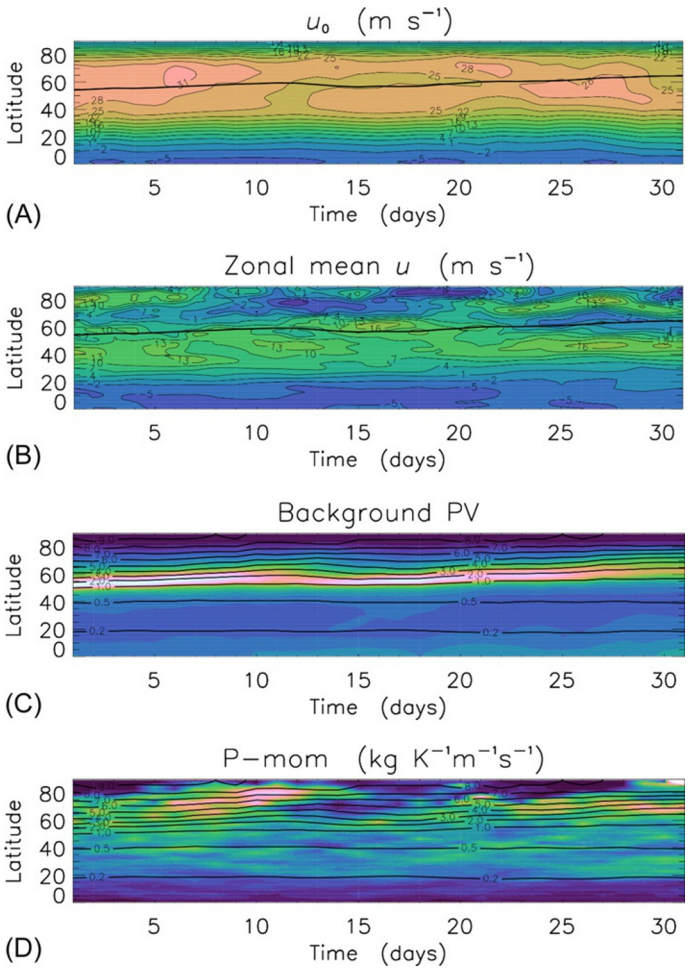

FIG. 1 Evolution of the atmosphere on the 320K isentropic surface calculated from daily ERA-Interim data for June 2007. (A) MLM background state zonal flow. (B) Eulerian zonal mean flow. (C) PV (PVU) in contours; color shading indicates the meridional PV gradient (the largest values are pink). (D) Wave activity (pseudomomentum density) in color shading overlain on background state PV contours (CI = 5000 up to 153,000 kg K^{-1} m^{-1} s^{-1} shown in lightest *pink*).

PVU $= 10^{-6}\,\mathrm{kg}^{-1}\,\mathrm{s}^{-1}\,\mathrm{K}\,\mathrm{m}^2$) on the 320K isentropic surface. It is obvious that it migrates poleward over the month, aside from a brief period in the middle. Fig. 1C shows the MLM state PV distribution on the 320K surface. All the midlatitude PV contours are migrating poleward, which can occur only through nonconservative (diabatic or frictional) processes. The meridional gradient of PV (color shading) also migrates with the tropopause on this surface. This is the seasonal march of the tropopause.

The mechanism is "vortex erosion" (Legras and Dritschel, 1993), where continuous filamentation of PV by the breaking of Rossby waves on the polar vortex edge transports mass within PV filaments away from the edge region into the surf zone, where the PV is mixed (McIntyre and Palmer, 1984). The net result is less mass within the vortex, and thus a poleward displacement. The high PV in the lower stratosphere (polar regions) is maintained by radiative cooling (Haynes, 2005), but this is weakest at the summer solstice; therefore, the high PV is not maintained. This occurs until late August, when the cooling begins to strengthen again and the high PV reservoir builds. The tropopause progresses slowly equatorward during the autumn. The time scale of delay behind the cycle in solar insolation, therefore, is related to the radiative equilibrium time scale of the troposphere (30 days, James, 1994). Fig. 1D shows a measure of wave activity on the 320K surface, which will be explained in the next section. It can be seen that the marked variations in the zonal mean are related to variations in wave activity, as argued earlier in this chapter.

2.2 Wave Activity Conservation Laws

Once we have isolated the background state, the partition is useful only if we also can anticipate the properties for evolution of the perturbations from it. The aim is to find a definition for wave activity that satisfies a conservation law of the following form:

$$\frac{\partial A}{\partial t} + \nabla.\underline{F} = D \tag{7}$$

which stems from the conservation laws obeyed by the full system combined with those satisfied by the background state due to its symmetry. In Eq. (7), A is a wave activity density, $_F$ is the wave activity flux, and D stands for the effects of nonconservative processes only. McIntyre and Shepherd (1987) set out a systematic approach to find the conservation laws for perturbations by combining globally conserved properties (such as angular momentum or energy) with properties, called *Casimirs*, which depend only on materially conserved properties. Two key examples are the mass and circulation enclosed by PV contours within isentropic layers, which can be described as functions of the PV and potential temperature (θ) coordinates.

The method proceeds by defining the pseudo(angular)momentum density as

$$P = -r(Z + S) + r_0(Z_0 + S_0) \tag{8}$$

where Z is the specific zonal angular momentum, $S(q,\theta)$ is the Casimir density (as yet unspecified), and the subscript zero refers to the same quantities in the background state. A central aspect is that there is a whole continuum of conserved properties, but the approach is to identify the property where the first-order contribution (in wave amplitude) is zero by construction, therefore ensuring that the resulting wave activity is second order (or higher).

Haynes (1988) showed the full nonlinear result for the primitive equations on the sphere which, in the limit of small wave slope, reduces to the more familiar form of wave activity density (see, e.g., Vallis, 2006, Section 7, for a simple derivation of the first term):

$$P = \frac{1}{2} r_0 Q_y \eta^2 - r' u' \cos\phi + \left(\frac{1}{2} {r_0}^2 q_0 {\eta_b}^2 - r_0 u' \eta_b \right)_b \cos\phi \, \frac{\partial \theta_{0b}}{\partial y} \tag{9}$$

where $Q_y = r_0 \cos\phi \, \partial q_0 / \partial y$ is the appropriate mass-weighted meridional PV gradient on the sphere, $y = a\phi$ is the meridional coordinate, and $\eta = -q'/(\partial q_0/\partial y)$ is defined as the meridional displacement of a PV contour relative to its latitude in the background state. Because the interior PV gradient of the background state is positive, the first term in Eq. (9) is positive definite; therefore, it is a useful measure of the amplitude of Rossby wave activity. Also, southward (negative) displacements give rise to positive PV anomalies through the advection of PV. The second pseudomomentum term, $-r'u' \cos\phi$, is often described as the *gravity wave term* because it is absent under quasi-geostrophic balanced dynamics and does not involve meridional PV fluxes. However, it can be an important player in some large-scale motions. For example, it is the dominant term in equatorial Kelvin wave activity.

The last term is proportional to the meridional gradient of potential temperature along the lower boundary. Because $\frac{1}{2} r_0 {\eta_b}^2$ is positive definite, the boundary term takes the sign of $r_0 q_0 \partial \theta_{0b} / \partial y$, which is typically negative (in both hemispheres) and opposes the interior term (although the terms involving u' are not sign definite). Note that η_b represents the meridional displacement of θ contours along the lower boundary in the full state relative to their position in the background state.

Fig. 2 shows the PV anomaly pattern (color shading) for a particular snapshot of the atmosphere on an isentropic surface (320K) that intersects the tropopause. The Ertel PV anomalies are weighted by the background-state isentropic density because this quantity, $r_0(q - q_0)$, reduces to the quasi-geostrophic PV anomaly field under the approximations of QG theory (see Section 12.4 of Hoskins and James, 2014). Because the density is much smaller at high latitudes in the stratosphere, this downweights the high-latitude negative anomalies; consequently, the positive anomalies are much more prominent. The positive anomalies are in tropopause troughs where the air has been displaced far equatorward and the tropopause is lower than its surroundings. Note that although each trough is distorted differently by advection, it is clear that there are seven centers of action around the midlatitudes. However, they are not equally spaced, with the strong anomalies over the eastern United States and Western Europe being separated the most, and the anomalies over Alaska and the west coast of North America being closest together.

Overall, the pattern projects most strongly onto zonal wave number 6 and the correspondence is highlighted by overplotting the wave number 6 Fourier component of the PV field. This serves as an indication that despite the nonlinearities introduced by advection and Rossby wave–breaking, the dynamics of the large-scale pattern may be interpreted in terms of wave propagation and interaction. This hypothesis will be tested using the ENM approach, which is derived by combining statistical analysis of data with wave-activity-conservation properties.

A similar, but less frequently used, conservation law exists for pseudoenergy. The large-amplitude derivation (Haynes, 1988) begins with a definition similar to Eq. (8), but using the

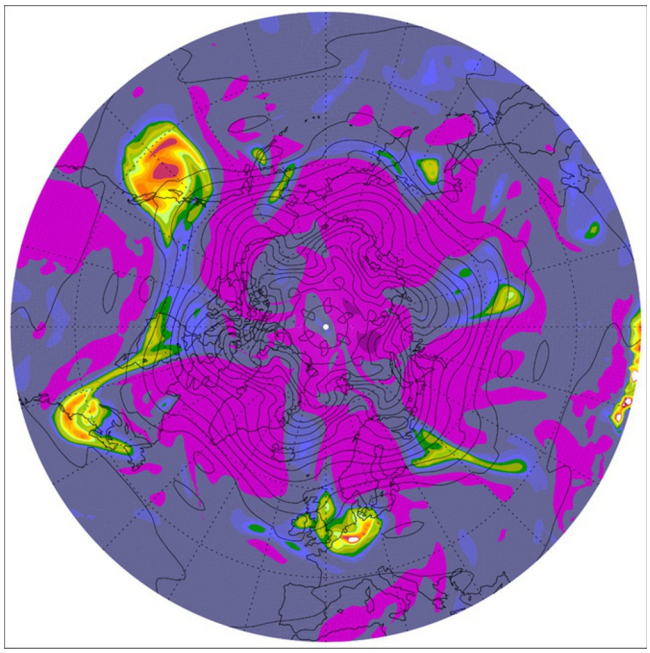

FIG. 2 Snapshot of the atmosphere (00UT June 23, 2007) on the 320K isentropic surface. The contour lines show PV, Fourier-filtered to zonal wave number 6 (interval 0.5 PVU), and the shading shows the PV anomaly, $r_0(q - q_0)$, from the unfiltered data overlain (negative anomalies in pink; the most positive anomalies in *red*; interval 10^{-5} s^{-1}). This pattern recurred throughout June and July 2007, and the trough (positive PV anomaly) over Western Europe gave rise to extreme monthly rainfall (Blackburn et al., 2008).

energy density in place of the zonal angular momentum. The pseudoenergy conservation law is then obtained using the time symmetry (i.e., steadiness) of the background state. Methven (2013) derived the small-amplitude expression for pseudoenergy, including perturbations near the lower boundary:

$$H = \frac{1}{2} r_0 \left(u'^2 + v'^2 \right) + \frac{1}{2} \frac{h_0}{g p_0 \theta} p'^2 - \frac{u_0}{\cos\phi} P \tag{10}$$

where the first term is the kinetic energy of the perturbations, the second is the available potential energy, and the third is called the *Doppler term* (for reasons which will become apparent), and is proportional to P (including the boundary terms). Note that perturbations are defined relative to the background state by using Eq. (1), where position is identified in isentropic coordinates (λ, ϕ, θ). For example, background pressure can differ from the full state at the same position in θ-coordinates.

Even at large amplitudes, the conservation laws imply certain properties for the perturbations. Consider a coherent disturbance that is neither growing nor decaying, but rather translating primarily along latitude circles. If the background can be defined as both zonally symmetric and steady, then the disturbance must have both a conserved pseudomomentum and pseudoenergy. The ratio of pseudoenergy to pseudomomentum gives the translation speed of the reference frame from which the disturbance appears steady to the observer. In other words, it defines the phase speed of the disturbance (Held, 1985; Zadra, 2000). This is the one of the central properties that we will use to characterize modes of variability in data.

2.3 The Implications of Wave-Activity Conservation for Modes of Variability

The small-amplitude limit of wave activity in Eq. (9) is precisely quadratic in disturbance amplitude, and this property has been exploited to make deductions about the disturbances in general and the properties of normal modes of atmospheric dynamics in particular (Held, 1985):

1. Because the wave activity is globally conserved, its rate of change is zero. Therefore, if the background state is also steady, disturbance amplitude can grow everywhere only if the global wave activity is identically zero. From Eq. (9), this leads to the celebrated Charney-Stern necessary condition for shear instability: The PV gradient must change sign somewhere within the domain. This argument includes the possibility that the negative pseudomomentum could be associated with the boundary wave activity in Eq. (9), as first described for baroclinic instability by Bretherton (1966).
2. Consequently, growing normal modes must have zero pseudomomentum, while neutral modes could have nonzero values.
3. Because normal modes evolve independently, each conserves pseudomomentum on its own. Therefore, if the global pseudomomentum of a superposition of modes is to be conserved, then the normal modes must be orthogonal with respect to pseudomomentum.
4. If the background state is also steady, pseudoenergy will be conserved and conclusions 2 and 3 also pertain to pseudoenergy. Also, because the disturbance energy is positive definite, from Eq. (10), we can obtain the Fjortoft necessary condition for shear instability: the zonal flow and meridional PV gradient must be positively correlated on average across the domain.
5. As argued in Section 2.2, the phase speed of neutral modes is given by the ratio of pseudoenergy to pseudomomentum. Considering the small-amplitude quadratic forms, we see that there are two distinct influences on phase speed:

$$c_p = -\frac{\langle H \rangle}{\langle P \rangle} = \left\langle \frac{u_0}{\cos\phi} P \right\rangle / \langle P \rangle - \langle E \rangle / \langle P \rangle \tag{11}$$

Here, the angle brackets denote an integral over the entire domain. The Doppler term in pseudoenergy gives the rate of advection of the disturbance by the zonal flow, even in the presence of shear. Because the wave frequency can be defined as $\omega = c_p k$, the Doppler term represents the shift in frequency associated with zonal advection by the background flow. Its sign depends only on the sign of the background zonal flow weighted by locations where the wave activity is largest. The second term describes propagation relative to the zonal flow and is proportional to the disturbance energy E. Because the energy is positive definite, the direction of propagation depends on the sign of the mode's pseudomomentum.

Note that it is not immediately obvious how to predict the phase speed of growing normal modes because they must have zero pseudomomentum and pseudoenergy. However, a solution to this problem was obtained by Heifetz et al. (2004), who used wave activity orthogonality to recast the growing and decaying normal modes obtained from complex conjugate normal mode solutions in terms of the linear superposition of a pair of counterpropagating Rossby waves (CRWs). By construction, CRWs are orthogonal with respect to

pseudomomentum, and therefore, one CRW has positive and the other CRV negative pseudomomentum such that their sum is zero when only the growing normal mode is present. Therefore, they describe disturbances that propagate in opposite directions (relative to the flow where wave activity is large). However, they are not orthogonal with respect to energy that can grow or decay as a CRW pair evolves.

They are used to give a mechanistic explanation for baroclinic or barotropic instability (a generalization of the Eady model for any unstable parallel zonal flow). The growth rate of normal modes can be expressed in terms of the energy of interaction between the CRWs. Perhaps most important, an expression for the phase speed of a growing normal mode is obtained as the average of the intrinsic phase speeds of the two CRWs:

$$c_{NM} = -\frac{1}{2}\left(\frac{\langle H_1\rangle}{\langle P_1\rangle} + \frac{\langle H_2\rangle}{\langle P_2\rangle}\right)$$

where $\langle P_1\rangle = -\langle P_2\rangle$ in the CRW construction.

3 THE ENM APPROACH TO OBSERVED DATA AND MODELS AND ITS RELEVANCE TO S2S DYNAMICS AND PREDICTABILITY

One way of diagnosing and characterizing the atmospheric S2S variability is to use a *phase space approach*, which has been shown to be very valuable in mathematics, physics, and atmospheric dynamics. The phase space of a geophysical fluid is a space where the state of the flow at a given time corresponds to one unique point. Usually, the phase space for a geophysical fluid can be represented by a steady basic state and superposed wave disturbances that are each represented individually as an oscillation in a 2D-phase plane (with characteristic amplitude and phase). In a nonlinear flow, this decomposition is nonunique (as discussed in Section 2.1), and the phase space trajectory can be complicated.

Here, we will focus on the evolution of waves and the insights that come from the small-amplitude limit for disturbances where linear wave theory applies. This is motivated by the prevalence of wave propagation on larger scales (e.g., Fig. 3), even though stirring by the large scales results in a continuous cascade of PV to smaller scales and the wavelike patterns associated with teleconnections. Even with such a drastic assumption, the proposed ENM diagnostic framework is very insightful and can be applied often to various type of flows with success, including nonlinear flow. Evidence from examples will be discussed next.

In general, waves transfer energy and momentum in flows through dynamical processes that obey the conservation laws introduced in Section 2. In Section 3.1, we will show that conservation laws constrain fundamentally the space-time characteristics of waves and are central to normal mode theory. For a given dissipative and stochastically forced flow, the conservation laws can be used to augment the physical relevance of the statistical principal component analysis (PCA) and its associated EOFs. In the context of waves relative to a steady basic state, if PCA is performed using wave-activity conservation laws, we demonstrate that the EOFs are the normal modes obtained from the linear wave theory.

The latter result permits the development of a statistical and empirical diagnostic framework with a built-in linear wave theory interpretation. It has been named ENM analysis by

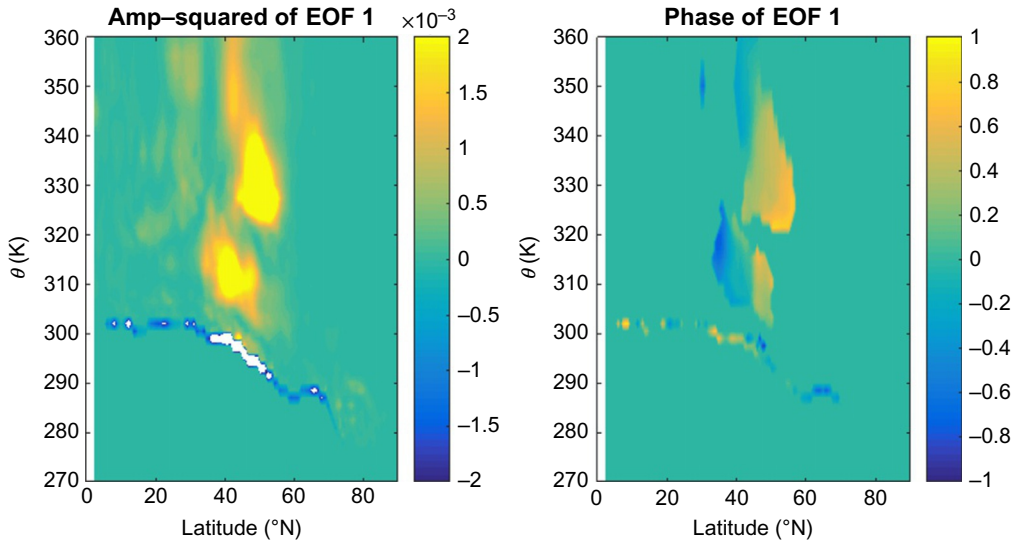

FIG. 3 The structure of the leading ENM pair for zonal wave number 6 during June and July 2007. (A) The amplitude squared of the ENM is its pseudomomentum structure, which is positive in the interior and negative in the boundary domain, which spans the space between the wavy and background state θ contours on the lower boundary (for each θ value). ENM is normalized such that the integral of amplitude squared over the latitude-θ section is unity (units $\text{rad}^{-1}\,\text{K}^{-1}$). (B) The phase of the meridional air parcel displacement associated with the ENM pair (radians/π).

Brunet (1994). In circumstances when the phase speeds obtained from ENM properties are consistent with the phase speed of waves observed by tracking (the ENM phase speed condition; see Eq. 19), we also will show in Section 3.1 that some aspects of ENM analysis are still conceptually and quantitatively relevant to stochastically forced and damped nonlinear flows. In Section 3.2, we will discuss the potential of ENM analysis for diverse applications, and in Section 3.3, we will focus on the ENM analysis of S2S variability.

3.1 ENMs: Bridging Principal Component, Normal Modes, and Conservation Laws

In Section 2.2, we discussed two well-known conservation laws for wave activity in geophysical fluid dynamics (Haynes, 1988). These wave activities are the total pseudoenergy for a steady basic state and the total pseudomomentum for a zonal symmetric basic state. When the basic-state wind is uniform, pseudoenergy and pseudomomentum reduce to total energy and total enstrophy, respectively, but shear in the basic state flow changes the dynamics fundamentally.

In the past, the quest for optimal bases has given rise to many independent rediscoveries of what is now known as *EOF space-time biorthogonal expansions.* A historical review of this topic is given by Sirovich and Everson (1992). The physical interpretation in terms of normal mode bases was explored by North (1984) for an atmospheric linear dynamical system with normal modes satisfying a self-adjoint equation. This work was extended by Brunet (1994) to normal

modes of the primitive atmospheric equation for sheared flows (in general not a self-adjoint problem) using the conserved pseudomomentum and pseudoenergy wave activities. In this more general case, the ENMs are not statistical eigenfunctions of a covariance matrix, but rather a solution of a generalized symmetric eigenvalue problem.

The statements about dynamical systems, normal modes, phase speed, and conservation laws discussed here can be found in Brunet (1994), Brunet and Vautard (1996), and Charron and Brunet (1999). For the rest of this discussion, we will assume that a zonal basic state exists, ensuring that pseudomomentum and pseudoenergy are conserved for an unforced and inviscid flow.

Consider a nonlinear dynamical system expressed in the following general form:

$$\frac{\partial X}{\partial t} = G(X) \tag{12}$$

where X is the state vector (e.g., the distribution of variables required to define the flow state), and $G(X)$ represents a dynamical operator that gives the rate of change of the state given the current state. If we linearize around a time-independent zonal basic state X_0 solution of Eq. (12), we obtain

$$\frac{\partial X'}{\partial t} = iG_0X' = iH_AAX' \tag{13}$$

where H_A and nonsingular A are time-independent Hermitian operators if and only if $W_A = \langle X', AX' \rangle$ is a conserved quantity (Charron and Brunet, 1999). This was demonstrated explicitly for the shallow-water model on the sphere for Rossby and gravity waves in Brunet and Vautard (1996). For the rest of this discussion, we will assume that W_A is the total pseudoenergy.

The bracketed term $\langle f, g \rangle = \int f^{\dagger} g dv$ represents an integral that can be over a line, area, or volume. The specific example is given by Eq. (10) for the dynamics described by the primitive equations on the sphere, where the state vector would need to be described by $X' = (u', v', p', r', \eta)$, including the perturbations in the interior and appearing in the lower-boundary terms, and the bracket represents the volume integral over the atmosphere.

Considering the normal mode expansion, in which each Z_n is a monochromatic wave solution of Eq. (13), then

$$X' = \sum_n a_n(t) Z_n = \sum_n a_{n,0} e^{i\omega_n t} Z_n$$

$$\omega_n Z_n = H_A A Z_n \text{ where } \quad \langle Z_n, AZ_m \rangle = \alpha_n \delta_{n,m} \text{ and } \quad \overline{a_n a_m^*} = \delta_{nm} \tag{14a}$$

in which α_n is the total pseudoenergy of the normal mode Z_n, a_n. ω_n is the natural frequency of the normal mode, and the overbar is the time average. This is a generalized eigenvalue problem, and because A and H_A are Hermitian operators and A is nonsingular, we can show for a space of finite dimension that the set of $\{Z_n\}$ forms a complete orthogonal basis under the pseudoenergy metric $M_A = \langle f, Ag \rangle$ (Bai et al., 2000). Hence, the normal mode expansion provides a full solution to the initial value problem associated with Eq. (13). The normal modes span a phase plane in phase space. They represent propagating waves except for stationary waves (i.e., $\omega_n = 0$).

If we linearize around a zonally symmetric basic-state X_0 solution of Eq. (12), then pseudomomentum $W_J = \langle X', JX' \rangle$ is conserved and normal mode solutions satisfy:

$$\omega_n Y_n = H_J J Y_n \text{ where } \langle Y_n, JY_m \rangle = \beta_n \delta_{n,m} \text{ and } \overline{a_n a_m^*} = \delta_{nm} \tag{14b}$$

where β_n is the total pseudomomentum of the normal mode Y_n. Note that the normal modes $\{Z_n\}$ and $\{Y_n\}$ are identical and form a unique basis if the eigenvalue pairs (α_n, β_n) are nondegenerate. We will assume for the rest of this discussion that this is the case. It follows from the framework of complete ensemble of commutating operators (CECO), where a set of many operators have the same eigenvectors if they all commute and their eigenvalues are determined uniquely for each eigenvector (Cohen-Tannoudji et al., 1973). For the primitive equations on the sphere with pseudoenergy and pseudomomentum conservation laws, we can write:

$W_A = cW_J$, and hence, for each normal mode, we have

$$c_n = \frac{\langle Z_n', AZ_n' \rangle}{\langle Z_n', JZ_n' \rangle} \tag{15}$$

where c is the mean phase speed of the flow (Held, 1985; Zadra, 2000), as discussed in Section 2.2 with respect to Eq. (11). This relationship also holds true for modified wave activities (by adding a divergent term that is conserved by construction) for individual isentropic surfaces that do not intersect with the surface (Zadra, 2000).

In the geophysical context, it is often more representative of observed flows to have a damped and stochastic forced model (for more on this, see Chapter 6). For the sake of simplicity, we will assume in the following that we have already decomposed Eq. (13) for each zonal wave number k.

Then the forced-dissipative version of Eq. (13) for each zonal wave number (ignore the subscript) can be expressed in the following form:

$$\frac{\partial X'}{\partial t} = iH_A A X' - \gamma X' + \varepsilon \tag{16}$$

where γ is a Raleigh-damping coefficient, and ε is a random forcing that is uncorrelated in time (e.g., Wiener process) and should be square-summable in space.

If we expand Eq. (16) in terms of normal modes $\{Z_n\}$, the complete time series is

$$X'(x,t) = \sum a_n(t) Z_n(x)$$

and the Fourier transform of the time domain in the equation yields the following coefficients:

$$\widetilde{a}_n = \frac{\widetilde{\varepsilon}_n}{i(\omega - \omega_n) + \gamma} \text{ where } \alpha_n \widetilde{\varepsilon}_n = \langle Z_n, A\widetilde{\varepsilon} \rangle \tag{17}$$

where the Fourier transformed variables, $\widetilde{g}(\omega) = (2\pi)^{-1} \frac{1}{T} \int_{-T/2}^{T/2} g(t) e^{-i\omega t} dt$ and α_n, are defined by Eq. (14a).

When $\gamma = 0$, a solution for Eq. (17) exists only if the Fredholm alternative (Riesz and Sz-Nagy, 1953) is satisfied; hence, $\widetilde{\varepsilon}_n|_{\omega=\omega_n} = 0$ for all n.

If the pair (α_n, β_n) as defined in Eqs. (14a), (14b) is nondegenerate, then in the limit $T \to \infty$ and using Eqs. (14a), (17), we can show that if A and B are nonsingular and we have a time series X' solution of the stochastically forced and damped dynamical system (Eq. 16), then for each zonal wave number k:

$$CJX_n = \beta_n X_n \text{ and } CAX_n = \alpha_n X_n \text{ with } \omega_n = k\frac{\alpha_n}{\beta_n} \tag{18}$$

where $\{X_n\}$ are the normal modes of Eq. (13) and the covariance matrix elements are defined by $C(x, x') = \overline{X'(x)X'(x')^*}$. A normal mode X_n obtained using a covariance matrix approach as in Eq. (18) is named an ENM.

The phase speed relationship (Eq. 15) implies that by knowing the X' time series, we have completely solved the initial value problem of Eq. (13), including its nonhomogeneous damped version with stochastic forcing (Eq. 16). This is possible because the unforced and nondissipative evolution equation (Eq. 13) is a completely integrable Hamiltonian system, which for a given truncation N has $2N$ constants of motion. Associated with the complete ENM basis $\{X_n\}$, we also have a complete orthonormal basis (i.e., principal components) in time $\{a_n\}$, where $\overline{a_n a_m^*} = \delta_{n,m}$. It means that each individual wave/ENM of a different type (e.g., gravity and Rossby waves) obtained through an ENM analysis will span biorthogonal subspaces in space and time with a clear dynamical interpretation.

ENM analysis permits, in practice, the ability to diagnose a specific wave spatial structure and its time evolution, which is not contaminated by other waves present in the flow. In particular, it will efficiently partition fast and slow modes without using any time-filtering technique. In Brunet and Vautard (1996), this has been shown to be very advantageous relative to a standard EOF analysis for simulated linear and nonlinear upper-tropospheric barotropic flows. It should be noted that other statistical techniques are available for studying and predicting atmospheric oscillations in the presence of damping and stochastic forcing. Two examples are the principal oscillation pattern (POP) method (Penland, 1989) and the constructed analog (CA) method (Van den Dool, 1994). They have been successfully used in long-range forecasting (e.g., Van den Dool and Barnston, 1995).

These two statistical methods rely fundamentally on temporal lag techniques for a given variable *time series* in which a linear *regression* equation is used to predict future values based on both the current variable values and the lagged (past period) values. It can be readily shown that, in general, these two are mathematically equivalent. The ENM analysis is fundamentally different because it is not based on time-lag correlation, but only on the existence of conserved wave activities. This makes ENM analysis very robust when dealing with noise. For example, the ENMs derived from Eq. (19) are not affected at all by a random reordering of the time series. This is not the case for POP and CA analyses because the covariance matrix depends only on time-averaging operations and hence is invariant under a reordering of the time series.

Of course, when performing an ENM analysis in practice, we need to assess the validity of the underlying normal mode assumptions for a given time series, such as the choice of basic state, conservation laws, and small-amplitude wave activities. The small-amplitude approximation can be relaxed by using finite-amplitude wave activities, but as discussed in Brunet (1994), the interpretation of the ENM analysis results is definitely more problematic. With the

exception of the research done by Brunet (1994), all ENM analyses to date have been performed using small-amplitude wave activities.

In general, we can objectively falsify the effectiveness of ENMs at representing the dynamic of atmospheric flows (whether simulated or observed). It can be done in the context of CECO theory and using temporal lag techniques, as in CA and POP analyses. One important aspect of such an evaluation methodology are the ENM phase speed conditions:

$$\overline{\Omega_n} = -\overline{i\frac{da}{dt}a^*} = k\frac{\langle X_n, AX_m\rangle}{\langle X_n, JX_m\rangle} = \omega_n \tag{19}$$

where $\overline{\Omega_n}$ are the observed mean natural frequencies. As discussed previously, these relations are necessary to demonstrate that an atmospheric flow is an integrable dynamical system. In practice, the principal components are generally not monochromatic, but the ENM phase speed conditions require that for each ENM, the observed mean natural frequency $\overline{\Omega_n}$ (derived from the principal component time series) is equal to its intrinsic natural frequency ω_n. These phase speed conditions have been verified within statistical estimation errors (e.g., dependence on the length of time series) for a wide variety of geophysical flows spanning mesoscale to planetary scales and have provided significant insights for many problems.

3.2 ENM in Applications Relevant to Predictability Across Time Scales

First, we will illustrate ENM analysis in application with a relative simple and explicit example taken from (Brunet and Vautard, 1996): the shallow-water model on the sphere in spherical coordinate. In many aspects, this model is relevant to the S2S prediction problem. It is a global barotropic model that supports Rossby, Rossby-gravity, Kelvin, and gravity waves typical of the midlatitude and tropical-upper-troposphere regions. Then for the ENM analysis of the evolution equation (Eq. 12), we have the following perturbation state vector for a given zonal wave number s:

$$X' = X - X_0 = \begin{pmatrix} u' \\ v' \\ \sigma' \\ P' \end{pmatrix} \text{ with basic state } X_0 = \begin{pmatrix} u_0 \\ 0 \\ \sigma_0 \\ P_0 \end{pmatrix} \tag{20}$$

where u', v', σ' and P' are the nondimensional zonal wind, meridional wind, height, and PV perturbations, respectively. The pseudoenergy A and pseudomomentum J operators are explicitly

$$A = \frac{1}{2}\begin{pmatrix} \sigma_0 & 0 & u_0 & 0 \\ 0 & \sigma_0 & 0 & 0 \\ u_0 & 0 & 1/F_R & 0 \\ 0 & 0 & 0 & -\dfrac{u_0\sigma_0^2}{\dfrac{d\mathrm{P}_0}{d\phi}} \end{pmatrix} \text{ and } J = \frac{\cos(\varphi)}{2}\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\dfrac{\sigma_0^2}{\dfrac{d\mathrm{P}_0}{d\phi}} \end{pmatrix} \tag{21}$$

where F_R is the Froude number and ϕ the latitude. Hence, from the considerations of the previous section, and with the sole knowledge of these operators, we can perform an ENM

analysis for a given shallow-water model time series X by solving the generalized eigenvalue problem (Eq. 18).

Note that the uniqueness and completeness of the ENM basis depend only on the rank of these two matrices (and of the covariance matrix). From their determinant we can show that this tantamount to have bounds a and j for which, for any latitude,

$$0 < \frac{\sigma_0^2}{\frac{d\mathrm{P}_0}{d\varphi}} < j \text{ and } 0 < -\frac{u_0\sigma_0^3}{\frac{d\mathrm{P}_0}{d\varphi}}\left(\frac{\sigma_0}{F_R} - u_0^2\right) < a \tag{22}$$

If they are not satisfied, the first and second conditions are related to the Charney-Stern and Fjortoft necessary conditions for shear instability, respectively, as discussed in Section 2.3. But the second condition also states that the wind should not match the local value of the gravity wave phase speed. The second stability criterion was first derived in a somewhat different manner by Ripa (1983), and it guarantees that there is no unstable normal mode. In the presence of unstable normal modes, the ENM analysis is still valid, but it needs a different approach, along the line explained for CRWs in Section 2.3. Of course, the unstable ENMs also can be obtained directly from the kernel of the pseudomomentum generalized eigenvalue problem (Eq. 18) (Martinez et al., 2010a).

Using the spectral shallow model on the sphere, Brunet and Vautard (1996), the ENM phase speed conditions (Eq. 19) were tested for linear and nonlinear regimes and various metrics (e.g., pseudomomentum versus the square of height) for typical Northern Hemisphere winter jets. The EOF diagnostics based on different metrics has clearly shown the advantage of using wave activities.

The time and zonal mean basic state of the time series was also the best choice. It minimizes the perturbation variance, and hence it is the closest to the small-amplitude limit. Of note, the phase speed conditions were satisfied for sheared, low-frequency Rossby waves in the linear regime only, for a relatively high spectral resolution of T100 compared to typical climate model resolutions in the 1990s (T32-64). The modes of variability of the shallow-water model with a realistic radius of deformation were well tuned only for sufficiently high spectral resolution. This example highlights the potential of ENM analysis as a suitable tool for assessing the numerical accuracy of dynamical processes in climate and prediction models. Zadra et al. (2002b) extended the method to multi-layer data from a primitive equation model and applied it to the Canadian Global Environmental Multiscale (GEM) NWP model dynamical core.

An important finding is that ENM phase speed conditions were verified even for nonlinear simulations with wave-breaking events (Brunet and Vautard, 1996). This indicates that over a sufficiently long period of time, the cumulative effect of nonlinear terms can be considered negligible when verifying the phase speed conditions, even if the ENMs are interacting nonlinearly. For example, although Rossby wave–breaking and the complexity of the flow results in the stretching and folding of individual troughs in different ways in Fig. 2, the propagation of the underlying wave pattern of PV anomalies is quantitatively predicted by Eq. (19). Therefore, ENMs are in some respects dynamically relevant to nonlinear problems. The framework for studying nonlinear interaction of normal modes in sheared flows and their relation to wave activities has been established in Vanneste and Vial (1994). This may

provide an avenue for studying the nonlinear interactions of ENMs with applications to the study of the predictability and dynamical processes of the S2S phase space and could pave the way for the empirical application of different chaos theory techniques to identify routes to chaos (e.g., KAM theory and period doubling).

In the literature so far, ENM phase speed conditions have been verified when analyzing global atmospheric North Hemisphere atmospheric variability on the 315K isentropic surface (Brunet, 1994), shallow-water-model Rossby waves (Brunet and Vautard, 1996), multi-layer NCEP winter reanalyses (Zadra et al., 2002a), simulated gravity waves in Charron and Brunet (1999), and for hurricane vortex, Rossby wave dynamics (Chen et al., 2003; Martinez et al., 2010a,b, 2011). In these studies, the ENM phase speed conditions were demonstrated within reasonable margins for a large portion of the studied wave activity variances.

Most recently, Methven et al. (2018) have extended the technique to the stratified problem, including the lower boundary, which introduces considerable complexity in using reanalysis data; but it is important due to the negative term introduced in the pseudomomentum and the corresponding term in pseudoenergy. In other words, the propagation of the potential temperature wave along the lower boundary modifies the phase speed of baroclinic waves, as is most familiar in the Eady and Charney models of baroclinic instability (Heifetz et al., 2004). Fig. 3 illustrates the structure of the largest amplitude ENM at zonal wave number 6 obtained from ERA-I data for June–July 2007. Note that in the calculation, the perturbations are defined relative to the MLM background state calculated by Methven and Berrisford (2015). The mode is propagating, and thus described by a pair of ENM structures in quadrature.

It has the largest amplitude in the upper troposphere and the lower stratosphere (straddling the tropopause) and a distinct negative pseudomomentum contribution associated with the boundary terms (here seen sloping along the lower boundary of the background state). Therefore, it is a baroclinic wave, although it has much stronger interior wave activity than the boundary term, and so it does not have zero pseudomomentum, as would be expected for a baroclinic growing normal mode. Its phase changes across the midlatitude jet in the troposphere, which is a signature of the different wave-breaking directions on either side of the jet. Evidence for this can be seen in Fig. 2 in the Fourier-filtered PV field. This particular structure meets the ENM phase speed conditions within a few meters per second (Methven et al., 2018). These studies clearly show the relevance of combining PCA, wave activities (including their associated Elliasen-Palm flux), and normal mode theory to diagnose atmospheric dynamics.

The advantage of the biorthogonal ENMs is evident: It permits a systematic approach for examining the statistics of data and exploring whether a modal perspective is dynamically relevant or not. The ENM phase speed conditions provide important nontrivial information on the dynamic processes (e.g., gravity versus Rossby waves) found in each of the subspaces spanned by the ENMs, as in hurricanes, where there is no separation of time scales between gravity and vortex Rossby waves due to a finite Rossby number (Chen et al., 2003). Similarly, the technique can be used to distinguish among types of wave modes when they are not separated, in terms of spatial scale. In this regard, it holds promise for addressing tropical dynamics in particular.

Falsification of the ENM phase speed conditions can happen for the following reasons: (1) nonlinear effects cannot be neglected; (2) the damping is not Raleigh or stochastic forcing is not Wiener; (3) the wave variability is not well simulated in the model; and (4) some dynamical and physical processes are not represented properly in the wave activities. The latter can

be sensitive to the basic state choice, as demonstrated in Brunet and Vautard (1996) with a shallow-water model. In some situations (e.g., baroclinic development), the important contributions of the boundary terms have been neglected (see Zadra et al., 2002a), as pointed out by Methven (2013).

The ENM approach can be used to study the response of a conservative dynamical system to arbitrary forcing. Nowadays, this is a key issue in climate science. According to Cooper and Haynes (2011), a study about the FDT, the response of a conservative dynamical system to steady forcing applied for $t > 0$, and in the limit, $t \to \infty$ would be equal in the ENM framework to $\delta x = -T^{-1}\delta f$, where T is simply the diagonal matrix composed of the intrinsic natural frequency of the ENMs. The advantage relative to the POP approach is that the response can be studied individually for each biorthogonal ENM with a direct physical interpretation. Of course, one of the focuses of climate studies should be on low-frequency dynamical processes like S2S variability.

3.3 ENM Application to the Atmospheric S2S Variability

In Brunet (1994) and Zadra et al. (2002a), large-scale atmospheric variability was shown to be spanned by ENMs, with intrinsic natural oscillation periods ranging from days to months. The ENM phase speed conditions were verified for almost all the wave activity except for some specific ENMs. For example, Brunet (1994) showed that the ENMs associated with Atlantic blocking did not satisfy the ENM phase speed conditions, possibly due to the nonlinear transient feedback and omission of boundary terms in the wave activities.

In Brunet (1994), the S2S North Hemisphere variability on the 315K isentropic surface for 24 winters was characterized quantitatively and empirically within an ENM framework. Fig. 4 shows the distribution of the total observed wave activity per day as a function of the ENM intrinsic natural oscillation period ($\omega_n = c_n k$) for various truncation thresholds. Eight discretelike ENMs were identified, which show a finite contribution to the total wave activity (over 1% individually), with distinct intrinsic periods from 14 to 200 days spanning the S2S time range. They represent around 20%–30% of the total wave activity and are closely linked to large-scale patterns like the AO and Atlantic blocking.

The rest of the wave activity (70%–80%) is represented by a continuous spectrum of ENMs with a peak around 3–5 days associated with transients, storms, and baroclinic waves (diagnosed as very distinct propagating ENM pairs, as shown in Fig. 3). The discretelike ENM pairs were shown to be relatively predictable, with a large e-folding time of 3–5 days. Here, the e-folding time is defined for each ENM pairs as the time average of its amplitude tendency (growth or decay) in the phase plane (Brunet, 1994), which is a good measure of predictability. The continuous spectrum has e-folding times of less than 3 days, which is consistent with predictability theory for baroclinic wave activity (Leith, 1978). The eight discretelike ENM pairs partly control the evolution of the continuous spectrum and the distribution of high-impact weather because they are large-scale features dominating the advection of PV. They are good candidates to span the phase space of a low-order model of S2S variability (see Chapter 6).

For example, Fig. 5 shows the phase-space-probability density of the zonal wave number 2 ENM pair, with an intrinsic period of 35 days for the winters from 1963 to 1987. An asymmetric bimodal signature with one localized small peak and one wide peak can be clearly

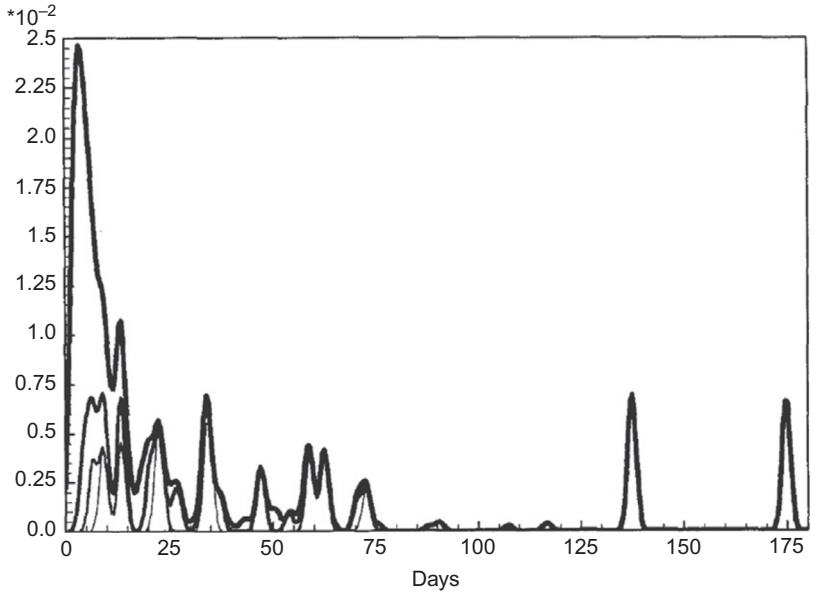

FIG. 4 Distribution of the observed wave-activity spectral density as a function of the ENM oscillation period for the NH 315K isentropic surface. The solid curves, from the thicker to the thinner, correspond to the total wave activity in percentage with individual ENM contribution higher than 0.1%, 0.4%, 0.7% and 1%, respectively. *From Brunet, G., 1994. Empirical normal mode analysis of atmospheric data. J. Atmos. Sci., 51, 932–952.*

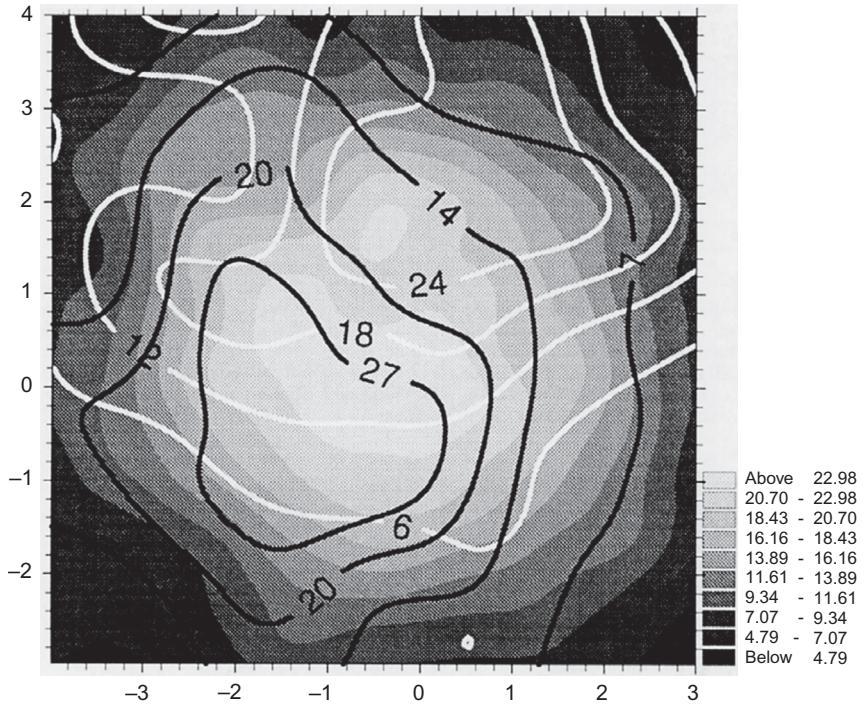

FIG. 5 Phase-space-probability density of the zonal wave number 2 ENM, with an intrinsic period of 35 days. The black and white contour lines correspond to the ZO and BL weather regime events, respectively. The probability density is shaded (units are 10^{-3}). *From Brunet, G., 1994. Empirical normal mode analysis of atmospheric data. J. Atmos. Sci., 51, 932–952.*

observed. A composite of wave activity maps shows that the small peak is associated with Atlantic blocking. This is confirmed in the same image by the probability densities of the zonal (ZO) and Atlantic blocking (BL) weather regime events obtained by Vautard (1990), where the ZO event density has a pattern similar to the main structure of the ENM bimodality.

It is noteworthy that the ZO and BL densities in Fig. 5 show a strong amplification of the ENM pair wave activity for the BL relative to ZO regimes, with an approximate π phase change in the amplitude that can be readily observed by noting that the BL maximum density is farther from the phase-plane origin than the ZO maximum density. This is observed for two other discretelike ENM pairs, and is typical of a resonant process. The presence of weakly unstable hemispherical normal modes in the slow variability is predicted by the wave-mean flow interaction theory of Charney and DeVore (1979); also see Chapter 6. They proposed orographically induced linear and nonlinear resonance mechanisms to explain phase locking and multiple equilibria of weather regimes.

It is recognized that the North Atlantic Oscillation (NAO) temporal variability spans many time scales and is subject to a wide range of atmospheric and oceanic forcing. For example, Molteni et al. (2015) has shown extratropical teleconnections with the Indo-Pacific region that have common atmospheric responses associated with different forcing spanning weeks to interdecadal time scales, which is consistent with resonant behavior. More in-depth studies of discretelike ENM phase space, weather regimes, and resonance mechanisms are needed to make progress on these S2S issues. One step toward this objective was achieved by the three-dimensional (3D) ENM analyses of the observed and simulated global atmosphere by Zadra et al. (2002a,b). The focus of Zadra et al. (2002a) was not S2S variability per se, but the upper-troposphere and lower-stratosphere observed variability along the tropopause. It has been shown that a large part of the wave activity around the tropopause was spanned by ENMs with intrinsic oscillation periods of less than 14 days, and this can be explained using the theory of quasi-modes (Rivest and Farrell, 1992).

Quasi-modes are defined as superpositions of singular modes that are sharply peaked in the phase speed domain but have large-scale (delocalized) structures, as opposed to singular modes, which represent sheared disturbances advected by the flow. They are often reminiscent of monochromatic discrete modes (neutral or unstable) that have been displaced into the continuous spectrum by the modification of the wind basic state (e.g., addition of wind shear) or dynamical processes like f-plane versus β-plane (Zadra, 2000). Quasi-modes are often weakly damped by critical layer stirring (Briggs et al., 1970; Schecter et al., 2000, 2002; Schecter and Montgomery, 2006; Martinez et al., 2010a) and maintain their energy for relatively long periods. They are easily excited by external forcing due to their large spatial structure, which is typical of discrete modes. They are not easily computed or identified by numerical or analytical methods (e.g., identification of relevant Landau poles), but they have been readily identified by ENM analyses in hurricane simulation (Martinez et al., 2010a) and atmospheric reanalysis (Zadra et al., 2002a) diagnostics. The latter study was able to identify leading modes with dipolar pressure patterns along the summer hemisphere tropopause that have well-defined phase speeds and decay rates of a few days, which can be explained by the theory of quasi-modes (e.g., wave number 5 with phase speed of 12 m/s and 3-day decay rate).

It is quite possible that the eight discretelike ENMs spanning the S2S variability in the 2D barotropic study of Brunet (1994) are quasi-modes. Further studies are required to confirm

this hypothesis, and these will probably need to be 3D ENM analyses because Zadra et al. (2002a,b) have identified a relatively larger number of discretelike ENMs with intrinsic periods of over 14 days spanning the S2S variability (e.g., more than a half-dozen for the zonal wave number 1 alone).

The ENM also could be used in the atmosphere-ocean S2S context where, in general, pseudoenergy is expected to be conserved, but not pseudomomentum. For example, the latter is not conserved when an oceanic basin has irregular boundaries because the zonal symmetry is broken. The ocean ENM diagnostic could be used to look at the MJO atmosphere-ocean coupled problem, with the ocean limited to the mixing layer. This would be the first step toward looking at ENMs of the ocean-atmosphere with sub-seasonal to multidecadal time scales. It is noteworthy that wave-activity study of the ocean is almost nonexistent except for the work of P. Ripa (Shepherd, 2003).

4 CONCLUSION

The first important objective of this chapter was to tackle the problem of diagnosing S2S atmospheric variability by splitting the diabatic and adiabatic flow components using fundamental principles from geophysical fluid dynamics. This was shown to be possible by taking advantage of the MLM theory based on the conservation of PV and potential temperature. The MLM partitioning of S2S variability in terms of slow diabatic processes, such as radiative forcing, and large adiabatic dynamical processes leads to the second important goal of our proposed methodology, which is to be able to extract dynamical modes of S2S variability from observed global data and model simulations with coherent space-time characteristics using fundamental properties that can be deduced from theory.

We demonstrate in this chapter that ENM analysis, with its built-in characteristics based on conservation laws, PCAs, and normal mode theory, provides an appropriate theoretical framework. ENM analysis is able to frame S2S scientific studies suitably and bring new perspectives (e.g., partitioning the S2S variability in fast and slow modes based on the ENM intrinsic phase speed). For example, the use of the ENM technique to date has revealed that a small number of structures dominate the observed variability at lower phase speeds than baroclinic waves. However, there are many unanswered questions regarding the nature of the modes. Are the ENM structures consistent with the structure of normal modes? Can we confirm or rule out their quasi-modal interpretation? Are they robust in a very long time series? The background state has a strong seasonal variation, but the calculation assumes a steady background, so should the modes be obtained for each season separately? What is the role of boundary wave activity?

Once the discretelike ENM phase space spanning the S2S variability has been characterized properly, we will be well positioned to address the S2S predictability problem. To address the predictability challenges, we need to understand the following points better:

- The extent to which a low number of distinct modes describe variability on S2S time scales and the robustness of those structures from year to year.
- How these discretelike modes interact with one another, with faster disturbances, and with the background state through nonlinear interactions. The role of physical mechanisms,

such as wave resonance, multiple equilibria, attractor sets, and stable and unstable limit cycles (see Chapter 6 for more on this topic), is also important.
- The role of the slow modes in predictability on S2S time scales.
- How the ENM approach can be used to examine tropical-extratropical interactions and the teleconnections that result in longer-term predictability in the extratropics.
- The degree to which ENMs present a reduction of atmospheric dynamics in terms of the average response of the system to stochastic and other forcings and use the phase space of ENMs to understand changes in S2S variability with climate change, including weather-regime responses to climate change (structure and occurrence), by performing intercomparisons of global and regional climate models as recommended by Palmer (1999).

Using statistical and theoretical research programs together will improve our knowledge of the S2S forecast problem and point the way toward exploiting new sources of predictability. For example, Brunet (1994) and the subsequent applications of ENM analysis to the observed and simulated global atmosphere (Zadra et al., 2002a,b) provided the guidance needed to look at the problem of the MJO and the NAO two-way interaction problem through teleconnections (Lin et al., 2009) and its impact on NWP skill (Lin et al., 2010a; Lin and Wu, 2011). As in the previous example, to make future advances in the S2S forecast problem, we also will need to use a hierarchy of GCMs of increasing complexity to gain the necessary dynamical and physical insights (e.g., Derome et al., 2005; Lin et al., 2007).

We believe that the S2S forecast problem is at the forefront of the weather and climate predictability continuum, where S2S variability can be represented by a finite number of relatively large-scale discretelike modes. These discretelike modes evolve in a complex manner through nonlinear interactions with themselves and transient eddies and weak dissipative processes. The sources of predictability are a mixture of fast adiabatic and slow diabatic processes that can be differentiated and diagnosed properly with a phase space approach based on ENM and MLM theories. Although the approach described here is not unique, the key to better prediction of S2S variability and weather regimes in a changing climate lies in improved understanding of the fundamental nature of S2S phase space structure and associated predictability arising from dynamical processes.

Acknowledgments

We thank Martin Charron, Yongsheng Chen, Yosvany Martinez, and Ayrton Zadra for their important contributions to ENM analyses. G. Brunet would like to express thanks to Hai Lin for his continuous leadership, discussions, and support throughout the years on the S2S diagnostic and prediction problem. Also, many thanks for the work of Tom Frame in cosupervision of the master's dissertation projects of Lina Boljka and Carlo Cafaro, which have advanced the ENM analysis, including the MLM background-state and lower-boundary-wave activity, and led to Fig. 3, and Paul Berrisford, for his contributions to the development of the background-state calculation. We thank Fréderic Vitart and Andrew Robertson for their helpful comments.