Variable Generation Power Forecasting as a Big Data Problem

Sue Ellen Haupt and Branko Kosović

Abstract— **To blend growing amounts of power from renewable resources into utility operations requires accurate forecasts. For both day ahead planning and real-time operations, the power from the wind and solar resources must be predicted based on real-time observations and a series of models that span the temporal and spatial scales of the problem, using the physical and dynamical knowledge as well as computational intelligence. Accurate prediction is a Big Data problem that requires disparate data, multiple models that are each applicable for a specific time frame, and application of computational intelligence techniques to successfully blend all of the model and observational information in real-time and deliver it to the decision makers at utilities and grid operators. This paper describes an example system that has been used for utility applications and how it has been configured to meet utility needs while addressing the Big Data issues.**

Index Terms— **big data, power forecasting, solar energy, variable generation, wind energy**

I. Introduction

Utilities and Independent System Operators (ISOs) depend on accurate forecasts for the next few hours to several weeks in order to effectively utilize variable generation resources. It is important to be able to forecast the wind, solar, and hydro power available the next day, or often, for the next several days. The marginal cost to run these renewable resources is quite low and it is economically advantageous to allocate as much power from those units as possible. But over-allocation of those units when the wind, irradiance, or water power is not available could lead to using much more expensive reserve units in real time. The specific rules depend on the particular ISO, but in general, the utility and ISO decision makers wish to allocate their resources a day or more ahead, and correct predictions of the power expected from the solar and wind units allows the marginal cost of energy to be minimized while assuring sufficient power to meet the load. Optimizing security of supply, economic dispatch, and power quality may often be conflicting demands, yet for all of these, having accurate forecasts of the renewable power is essential [1]. In real-time, grid operators must have very short-range forecasts (known as nowcasts) to meet energy demand and to minimize the cost of running excessive reserves. Fig. 1 shows an example of the differences between power output by a solar unit on a clear day vs. a day with variable clouds. It is obvious that large variability in output exists. Wind power also exhibits considerable variability. Some system operators (such as the Electric Reliability Council of Texas – ERCOT) use historical error statistics to help determine nonspinning reserve requirements while others (such as the Sacramento Municipal Utility District – SMUD) is interested in spatial and temporal variabillty of solar irradiance for energy trading and generator dispatch decisions [2].

To forecast across scales from seconds to a few weeks, forecast methods must be tailored for each scale [1]. We focus here on the short-term (from times 0 to about 6 hrs, used for pre-dispatch and scheduling small power systems) and the medium term (current to about a week out, which are used for pre-dispatch, unit commitment, trading, and maintenance planning) [1]. Observation-based nowcasting provides a much more accurate forecast in the short range, but its skill drops off rapidly with time. Numerical weather prediction (NWP) becomes more important at about 3 hours and provides value to about 2 weeks. The limits of predictability of NWP are currently around 10-14 days. Because NWP simulations at high resolution over a sizable domain can require the order of hours to run on supercomputers and often require spin-up time, it is not typically available for real-time use for the shortest time ranges.

Modern methods of forecasting renewable energy employ post-processing methods to blend these disparate models, as well as ensembles of model runs, which greatly improve the forecast skill [1-7]. Ensembles of model runs also provide probabilistic forecasts. Forecasts of the wind speed expected at turbines and irradiance at solar panels can be converted, using machine learning, to power forecasts that meet the needs of the utilities and ISOs.

Submitted 28 April. This work was supported by the U.S. Department of Energy under award number DE-EE0006016.

S.E. Haupt and B. Kosovic are with the National Center for Atmospheric Science, Research Applications Laboratory, Boulder, CO 80303 USA (e-mail: haupt@ucar.edu, branko@ucar.edu).

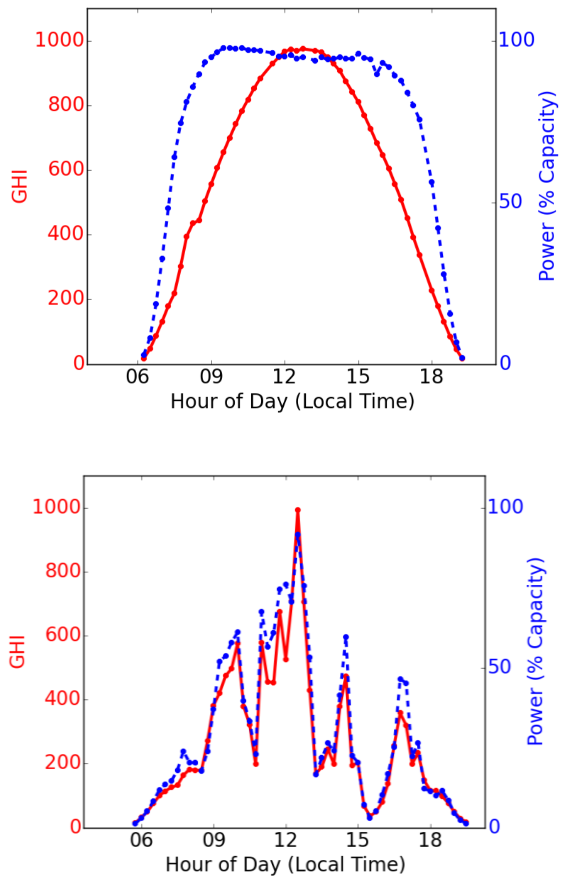

Fig. 1. Energy output from a commercial solar plant on a clear day top) and from the same plant on a cloudy day (bottom) The solid red curves show the global horizontal irradiance (GHI) and the dashed blue curves that power output from the commercial array as a percent of its capacity.

The methods to make the forecast require a plethora of historical and real-time data from both models and observations, blended in real-time to provide a seamless forecast for use by decision makers. Thus, forecasting for variable renewable energy, namely wind and solar power, is an example of a real-world Big Data problem and is best treated as such [8]. Here we refer to Big Data as including collections of datasets that are large or complex and where special considerations are necessary for processing to reveal patterns and extract actionable information. Characteristics of Big Data are often stated to include its volume, velocity, variety, variability, veracity, and complexity (https://en.wikipedia.org/wiki/Big_data). Each of these is discussed in detail below.

This paper lays out the issues for forecasting the weather for renewable energy in Section II. Section III discusses weather forecasting as a Big Data application and how fits the characteristics listed above. Section IV provides a case study of the Sun4Cast™ solar power forecasting system designed by the National Center for Atmospheric Research (NCAR). The final section V summarizes the issues and presents prospects for future applications.

II. The Necessities of Weather Forecasts

Weather forecasting has always been one of the original computational challenges. From the time that L.F. Richardson imagined a room of human "computers" numerically solving the primitive equations of fluid mechanics [9] in 1922, meteorologists have been seeking to use data mining and the best numerical methods to improve forecasts. A filtered version of these equations was the first problem on the first operational computer, the Electronic Numerical Integrator and Computer (ENIAC) at Aberdeen Proving Grounds in 1950, set up by Jules Charney, John von Neumann, and R. Fjortoft [10]. This was the beginning of meteorologists' passion for using computing to make advances in forecasting that naturally leads to consumption and production of Big Data. The detailed physics included in the numerical models, as well as the spatial and temporal resolution have been enhanced rapidly, which constantly challenges computing capability. At the same time the field has quickly adopted advanced statistical and computational intelligence methods.

Weather prediction is an important real-time challenge and many applications rely on its accuracy, including energy applications, aviation safety, defense planning, and many more. To meet those expectations, NWP models are run as often as hourly at high resolution over tens of millions of grid cells and include physics packages that solve their own series of equations. These physics packages model or parameterize incoming and outgoing radiation, cloud physics, shallow and deep convection, boundary layer turbulence, land surface interaction with the fluid atmosphere, and more.

The equations of atmospheric motion are nonlinear and dissipative, which make them chaotic, implying sensitivity to initial conditions. To deal with this chaotic atmospheric flow, two approaches are often adopted. First, data assimilation blends observed data into the initial model state. These data come from sampling the horizontal and vertical extent of the atmosphere as well as the land and sea surface boundaries. Some of these data are remotely sensed either from the ground, such as by radar, or from satellites. Thus, the data points are seldom located on the grid and are often of disparate nature. A second approach to dealing with the chaotic nature of the flow recognizes the potential for multiple possible realizations of the development of the weather event. This approach simulates these realizations as an ensemble comprised of many runs with slightly different initial or boundary conditions or different physics parameterizations. Some centers run upward of 50 model ensemble members to form a probability density function (PDF) of the development of the weather [11].

Finally, the best predictions are post-processed to blend as much of the data, model output, and statistical learning as possible to improve the deterministic forecast and to quantify the uncertainty. As model output regularly includes systematic (bias) and random errors, additional observations, when available, can be used in the post-processing step to improve the forecast. Training these post-processing methods requires large amounts of both model and observational data as discussed in more detail below. The best methods blend computational intelligence with the discretization of the physics and dynamics of the system. Such systems can be quite complex [3,12} and is certainly a problem in Big Data [8].

III. Weather Forecasting for Solar Power as a Big Data Problem

Modern applications of weather forecasting include all characteristics of Big Data include its volume, variety, velocity, variability, veracity, and complexity. How each of these applies is described herein.

A. Data Volume

Producing real-time NWP model simulations is a notorious challenge for large-scale computing. NWP models ingest and assimilate large amounts of observational data to initialize a model run. They then solve the nonlinear Navier-Stokes equations on grids upwards of 100 million grid cells with time steps on the order of 20 s. This implies that roughly 18 billion calculations of each model variable are handled for each hourly output. Thus, these computations require large scale computing resources. When one multiplies the degrees of freedom of the problem (grid size times dozens of variables times the time increment) by 168 hours for each week of forecasts, one begins to appreciate the large volume of data (hundreds of trillions) in the calculations. Of course not all of that data is stored, with about 30 output variables being stored at 15 min increments and another 700 stored at hourly increments. The national centers require large supercomputers to produce these simulations. Storage is optimized by archiving only those data that will be subsequently used.

B. Data Variety

To produce a forecast requires combining various types of observations, some of which sample surface variables at convenient locations that seldom correspond with the NWP grid points, while others represent the vertical profile of the atmosphere at specific locations. Remote sensing, such as from a satellite, sample other variables on a horizontal grid at differing elevations. Satellite irradiance values often depict the cloud top temperature. Since different clouds appear at differing atmospheric levels, these indicate temperature at different levels of the atmosphere. Other satellite instruments, however, look through the clouds to determine properties of the atmosphere in vertical profiles. Other remote sensing instruments, such as surface-based radars, scan the environment and provide reflectivity in terms of distance, angle, and azimuth of the beam.

Even though these remotely sensed observational data may be gridded, they are not necessarily on the same grid, or even use the same map projection as the NWP data. Thus, one must include interpolation as a necessary step of any forecast process, for both point-based observations as well as the remotely sensed data. Moreover, all of these differing types of data with differing types of grid systems must be coordinated to provide a picture of the current atmospheric state before it is used to initialize the forecasting system.

The existence of standards and standardized formats for meteorological data, including metadata, significantly reduces the possibility of errors when processing these disparate data. Unfortunately, no such standards exist at present for data collected at power plants, although standardization is essential to develop accurate and robust renewable energy forecasts.

Some common problems occur with the lack of power plant data standardization. One common standardization issue is the time stamp for the observations. All observations and NWP data should include a time stamp in universal time. However, that is not always the case with power plant data. In particular, it is common for specialized observations to be listed in local time. For such observations, it can be confusing to determine the time zone and whether or not the reported local time is in standard or daylight saving time, as some systems report local standard time year-round, and some switch with the twice-yearly time changes.

Another common data standardization issue lies in the averaging time of the observations. There are many small networks of specialized weather observations (mesonets). Although some of these are standardized, not all follow standardization on reporting details and averaging periods. For instance, one dataset that brings together observations from a variety of mesonets includes data with differing averaging periods, with hourly data including everything from averages of 10 Hz data for the full hour, for 15 min, for 5 min, for 1 min, and even for an instantaneous value. Some observations are recorded at the top of the hour, while others are an average of the prior hour. It is often difficult to standardize the values that are provided. It is critical to understand the metadata describing issues such as averaging period when developing best methods to deal with the frequent disparity. These challenges must be met by each group using the data, but efficiency could be gained through standardization.

C. Data Velocity

Having large amounts of data arrive at different times creates a significant challenge for processing. One must prepare for different arrival times for each NWP model and each observational dataset. This implies that as data arrive, they must be matched to the valid time and steps taken to account for any lags before blending it with data from other models, observations, or systems. Systems such as those built at NCAR incorporate these data as they arrive. Thus, system engineers must take into account and track the arrival time when doing the integration.

D. Data Variability

With the data acquisition speed so variable, one must expect that it will be common for some of the data sources to be delayed. Thus, one must plan for graceful degradation of predictions when particular sources of data or model output are not available in time to provide the real-time prediction. Because this is a frequent occurrence, fall-back routines are necessary for each type of data that could be missing. Although the computational intelligence algorithms are trained to optimize on having all of the data, it is necessary to also provide forecast model systems that assume that some of the data are missing. This process becomes yet more complex when more than one data source is missing. Again, this requires good software engineering preparation to handle the contingencies.

E. Data Veracity

The quality of the data is a critical issue for both training the computational intelligence models as well as for real-time calculations. One must be prepared to identify issues with incorrect data that must be corrected. For instance, when an observation is far from the expected range for the season and time of day at a location, it can be flagged for potential error. An additional check on the previous value of temperature can determine if the change in that time period is within reason. One must take into account, however, that occasionally, rapid changes or anomalous values of weather variables may be real. Rapid temperature changes do occur, such as during passage of a weather front. In addition, extreme values of weather variables also occur. During times of flooding, for instance, precipitation observations could appear anomalous when in fact they are correct for that unusual case. Thus, it is important to construct quality control algorithms to identify these possibilities.

F. Data Complexity

A final characteristic of Big Data is complexity. The previous description of the issues of data volume, variety, velocity, variability, and veracity exemplify the complexities of attempting to blend these data to provide accurate forecasts in real-time.

For example, there are several uses of observations in making the forecast: 1) assimilation into the NWP models, 2) training the computational intelligence algorithms, 3) using historical data corresponding to a past prediction that is analogous to the current model prediction to predict the future state (as in AnEn approach described below) and 4) identifying the current conditions. Data are again required for verification and validation after the prediction is made. Thus, it can be challenging to take the best advantage of the data for each purpose without compromising the other uses.

IV. THE SUN4CAST™ SOLAR POWER FORECASTING SYSTEM

Solar power forecasting provides an example of applied weather forecasting where correctly modeling the variability is an important goal. Here we describe the Sun4Cast™ solar power forecasting system, which is a new comprehensive approach to forecasting the power produced from the sun's irradiance and includes a variety of components that illustrate the difficulties inherent in making such forecasts and directly deals with those difficulties. NCAR has worked closely with utilities and ISOs to produce forecasts that allow them to effectively balance the variable generation resources with conventional energy sources. In order to meet both the short-range (nowcast) and longer-range (day ahead and beyond) needs, NCAR forecasts the expected irradiance and the resulting power output from 15 min through 168 h. This system was recently built through a Public-Private-Academic Partnership funded by the U.S. Department of Energy (DOE) to advance solar power forecasting. The project sought to advance the state-of-the-science to improve irradiance and power forecasting and involved several utilities as cost share partners. To accomplish the goals, one must be able to forecast the aerosols and clouds accurately. Forecasting clouds, in particular, has proven to be challenging in the past. This comprehensive system was designed with both the needs of the intraday unit commitment and dispatch, as well as longer-range unit scheduling and planning in mind. The architecture is summarized in Fig. 2.

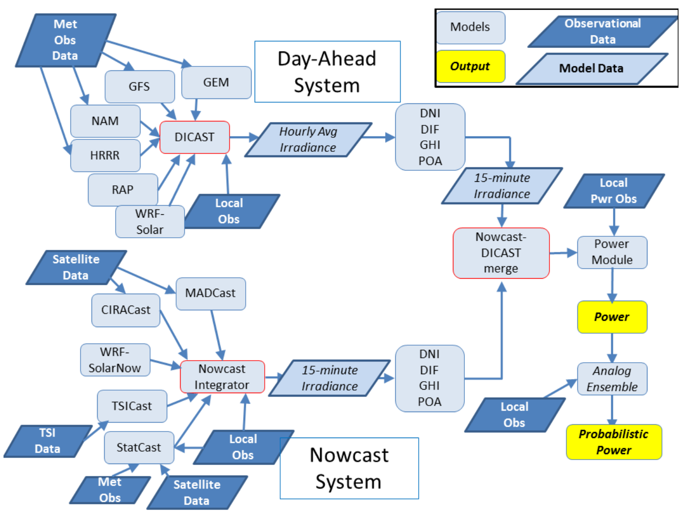

Fig. 2. Diagram of data flow in the Sun4Cast solar power forecasting system.

A. NWP Forecasts

For the forecast times beyond the nowcasting period (beyond about 4 hours), Sun4Cast leverages NWP models run by the National Centers for Environmental Prediction (NCEP) of the National Oceanographic and Atmospheric Administration (NOAA) and other national centers, as well as deploying NCAR's WRF-Solar. Each of these models has its own grid and timeframe.

1) Specialized runs such as WRF-Solar™

WRF-Solar™ is a newly developed branch of the Weather Research and Forecasting (WRF) model designed specifically to improve solar irradiance forecasts. This version includes an improved radiative transfer scheme, improved cloud physics parameterization, new shallow convection scheme, improved equation of time, robust interaction between the clouds and aerosols with the radiation, and output tailored to the specific application. Initial and boundary conditions for the WRF-Solar forecasting system derive from the Rapid Refresh (RAP) model analysis. The irradiances (global horizontal irradiance - GHI, direct normal irradiance - DNI, and diffuse irradiance - DIF) are output every model time step (20 s) and one-minute averages are computed [22]. It is run with one primary domain of 3-km horizontal grid spacing over the US and two domains of 1-km grid spacing over regions with solar farms – the San Luis Valley in Colorado and Sacramento, California. One run per day is configured to meet operational needs of the private partners targeting the day-ahead forecast. The model is initialized at 0000 UTC and is run for 54 hours in order to provide a day-ahead forecast available at the beginning of the morning. The model therefore runs during the night period so that the latency of the system is not relevant. Because the computational cost of activating the 1-km domains is high, they are only activated for the daytime of the second day of the simulation to enable the simulation to complete in time for the forecast.

2) HRRR

NCEP's High Resolution Rapid Refresh (HRRR) model is an example of a model that forecasts the weather over a limited area domain, in this case the CONUS, using a relatively fine 3-km grid cell size. It produces forecasts hourly for the next 15 hours. A single forecast consisting of three-dimensional fields amounts to 365 Mb, while corresponding two-dimensional fields representing surface conditions account for an additional 84 Mb, for a total of approximately 450 Mb for each simulation.

3) RAP

The RAP model, version 2 includes a wider domain than the HRRR. It is run hourly with a coarser grid to produce forecasts for the following 18 hours. Each RAP output file amounts to 55-60 Mb.

4) GFS as a typical global model

NCEP's Global Forecast System (GFS) is representative of the available global models. GFS is run at 2.5°, 1.0°, and 0.5° globally. The model forecasts are produced every 6 hours out to 384 hours. Recently an additional 0.25° simulation was added and produces forecasts for the next 168 hours. The Sun4Cast forecasting system employs the 0.5° forecast, using 65-70 Mb of data. Other global models, such as those from Canada, Europe, and other national centers can also be blended.

Table 1 summarizes the data output by some of the NWP models used in the Sun4Cast system.

TABLE 1
DETAILS OF SEVERAL NWP MODEL DAILY OUTPUTS.

Model	Forecast frequency	Hours ahead	Grid cell size	Daily output [GB]
HRRR	hourly	15	3 km	130
RAP	hourly	18	9 km	5.7
NAM	6 hours	84	12 km	5.5
GFS	6 hours	384	0.5°	68
GEM	12 hours	240	1°	4.3
WRF-Solar	Irradiance only: 20 sec	30	3 km (CONUS) 1 km (2 subdomains)	4.2

B. Nowcast System

Five models comprise the Nowcast system and each displays a "sweet spot" for producing a most accurate forecast. These nowcasting methods leverage a variety of disparate observational data, statistical and computational intelligence methods, and physical understanding of the atmosphere to produce a "best practices" blended forecast. Each is briefly described below.

1) TSICast, built and deployed by Brookhaven National Laboratory, uses three total sky imager (TSI) cameras to observe current cloud cover. Because they deploy multiple cameras, they can deduce the height, base, location of the clouds, as well as the speed and direction of each cloud layer by observing the changes in time. Thus, they can predict where the clouds will be in the next 15-30 min [13]. TSICast processes in about 2-3 min to provide this short-range prediction.

2) StatCast was developed by NCAR and Penn State University to leverage irradiance measurements from pyranometers located at the solar plant. There are several versions of StatCast [14-16], each of which uses a computational intelligence method to predict the cloud cover and the resulting clearness index for the next 3 hours. It ingests surface irradiance measurements, nearby weather data, and, when available, satellite data to estimate the clearness index (the observed surface irradiance divided by that available at the top of the atmosphere at that location). StatCast requires at least a year worth of data to train the forecast model. Once trained, it runs in a matter of seconds.

3) CIRACast was designed by Colorado State University's Cooperative Institute for Research in the Atmosphere (CIRA) to detect geostationary satellite-observed clouds, process the data to remove parallax and shadowing, and advect those clouds with derived motion vector and model winds [17]. Thus, they are able to predict cloud coverage over the coming hours. Its latency depends on the time to process and ingest remotely the satellite and model wind data, typically around 15-30 min. This is still useful for the next several hours as it provides the best "big picture" view of the state of clouds and can be used to project their location for the next several hours. It has been particularly useful for predicting short-range ramps in solar power.

4) The Multi-sensor Advection Diffusion foreCast (MADCast) system uses the Multivariate Minimum Residual (MMR) scheme of Auligné [18-20] to assimilate satellite infrared radiance observations into the dynamic core of the Weather Research and Forecasting (WRF) model. The dynamics of WRF then advects the observed clouds

accordingly. It predicts out to 6 hours, with a latency of only about 10 min due to not employing the computationally expensive physics packages of WRF.

5) WRF-Solar-Now is an implementation of the specially configured version of the WRF model, WRF-Solar that optimizes computation of solar irradiance (see details below). It is run in a nowcasting mode at 9-km horizontal grid spacing over the contiguous United States (CONUS) hourly. It predicts out to 6 hours with approximately 1 h of latency to complete the run.

The Nowcast system has different data needs, most of which are more modest than for NWP. The amount of data produced by the Nowcast system is displayed in Table 2.

Observations used in the Nowcast system include irradiance, air temperature, and power output. The total amount of data received daily for all the sites for which forecasts are produced (14 Sacramento Municipal Utility District, 2 Xcel Energy, 9 SoCal Edison, and 25 Brookhaven National Laboratory) is approximately 35 MB. While this amount of data is modest, data quality and disparate data formats represent a processing challenge. An exception to the modest data size is the satellite data. For CIRACast, approximately 1.54 GB/day in raw GOES data (GOES-W and GOES-E) is pulled from the satellite feeds plus 1.9 GB/day for the PATMOS-x L2 data uses for the advection forecast. In addition, they use the GFS model data (about 80 GB of data daily, of which 2-3 GB is directly applicable to the CIRACast forecast.

TABLE 2
DETAILS OF SEVERAL NOWCAST MODEL DAILY OUTPUT

Model	Forecast frequency [minutes]	Hours ahead	Daily output [MB]
MADCast	15	6	2,100
CIRACast	15	6	1.4
StatCast	15	3	13
WRF-Solar-NOW	15	6	24,000

C. Completing the Forecast

The integrator for the various NWP models is the computational intelligence algorithm, the Dynamic Integrated Forecast System (DICast®) [23], as depicted in Fig. 3. DICast® produces automated forecasts using a method that was designed to emulate the human forecast process. It generates forecasts by optimizing the combination of NWP model data through developing empirical relationships gleaned from historical model output and observations. DICast typically reduces root mean square error by about 10-15% and essentially eliminates bias as compared to the best input model. DICast employs a two-step process: it first statistically corrects the bias of each input model using Dynamic Model Output Statistics (DMOS) [24], and second, it optimizes the model blending weights for each lead time, producing a consensus forecast. DICast typically works with up to 90 days of data; this is an advantage because many other methods require a year or more of data for training, during which time some of the models may have been modified or upgraded, making the training process difficult.

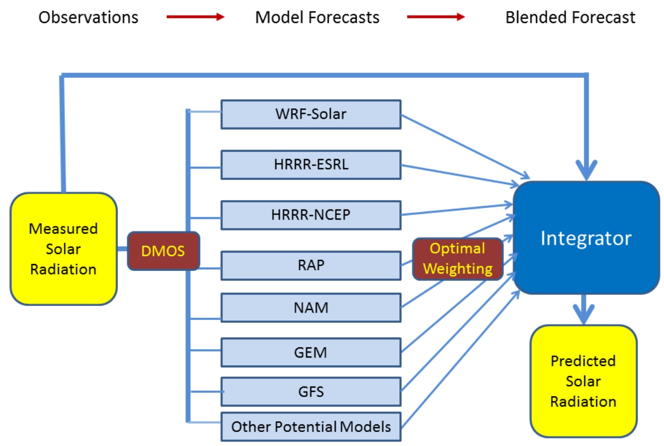

Fig. 3. Diagram of the DICast® blending process.

The configuration of DICast used in Sun4Cast employs irradiance values obtained from seven NWP models as seen in Fig. 3, as well as observations from the sites of the solar plants. These models include those run operationally by NCEP in the USA: Global Forecast System (GFS), North American Model (NAM), Rapid Update model (RAP), and High Resolution Rapid Refresh (HRRR-NCEP) as well as the HRRR-ESRL research model developed at the Earth System Research Laboratory of NOAA. Environment Canada runs the Global Environmental Mesoscale (GEM) model. WRF-Solar was run quasi-operationally by NCAR [25]. The data from each of these models is ingested and blended in real time to produce forecasts hourly.

The Nowcast systems discussed above are integrated separately using a unique Nowcast expert system integrator that utilizes the recent performance scores of each component model, whether it be from a computational intelligence method (StatCast), based on cloud observations (TSICast and CIRACast), or includes NWP components (MADCast and WRF-Solar-Now). Although the Nowcast system is currently optimized via an expert system, dynamic methods are planned for future applications. Each of these models has been shown to provide value in the system [25]. Fig. 4 depicts the mean absolute error of each of the models for each lead time out to 6 hrs. We see that the models each show value over some time period and all outperform the baseline persistence model (black line) over all except the shortest time periods.

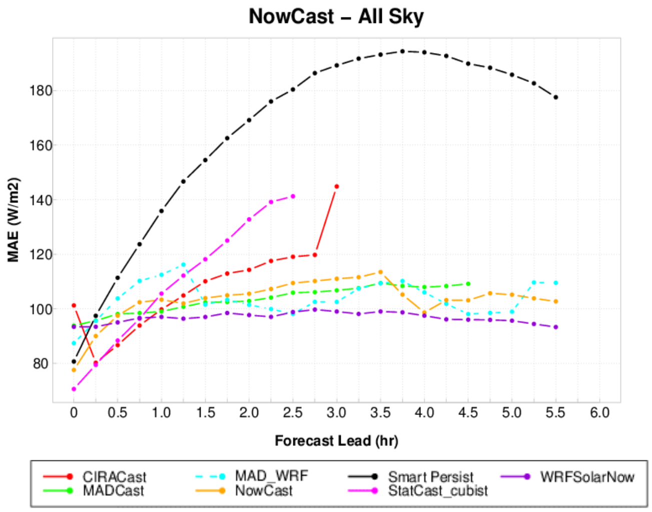

Fig. 4. Mean absolute error in W m^{-2} for all NowCast components aggregated over all partner sites and all sky conditions.

The DICast® and Nowcast irradiance forecasts are integrated and blended during the transition period (2 hrs - 6 hrs) to produce irradiance forecasts for each 15-min interval out to 3 hours then hourly out to 168 hours. GHI is the most useful forecast variable for photovoltaic panel operations (see Fig. 1) while DNI is the only component useful for concentrated solar plants. Diffuse irradiance DIF relates the two. Not all of the NWP models, however, are able to separately forecast DNI and DIF. That was one of the advances made when formulating WRF-Solar.

The meteorological irradiance values are not the final output variables. Utilities require a power prediction, meaning that an irradiance-to-power conversion step must be added. A model regression tree (Cubist) is used in Sun4Cast to train the relationship between the measured irradiance value and the coincident power produced. The empirically derived relationship is then applied in real-time to the irradiance forecast to produce a power forecast. A separate power conversion algorithm must be trained for each generation site. Once the training/testing procedure is accomplished, the algorithms' real-time application runs in a matter of seconds.

Finally probabilistic forecasts have been requested by partner utilities. NCAR applies the Analog Ensemble (AnEn) approach [26,27] to produce an appropriate probability density function (PDF) of the forecast uncertainty. The AnEn assumes that if a forecast made in the past under meteorological conditions analogous to today's forecast, then it is likely to produce the same error characteristics as is probable in today's forecast. Thus, analogs in those past forecasts are identified so that: 1) observations corresponding to analog forecasts are selected as members of AnEn and used to correct the forecast, and 2) a PDF of multiple analogs is used to estimate the uncertainty of the forecast. This flow-dependent uncertainty has been shown to reproduce the forecast and its statistical reliability at least as well as the full ensembles of runs produced at the operational centers [26,27].

NCAR produced probabilistic solar power forecasts in quasi-operational mode hourly and made them available to the utility and ISO partners for a year in order to provide sufficient data for a full assessment. Results indicate that each component improves upon baseline forecasts and has a "sweet spot" where that component often produces the best forecast and contributes to an improved forecast [25]. Over the three years of the project, the overall forecast accuracy improved by roughly 50% [25].

C. *Forecast Usage*

The Sun4Cast forecasts were available to the utility and ISO partners to use in integrating their solar resources into their energy mix. The partners use those forecasts to plan their day-to-day operations. One use is in allocating their units for the next day. Early in the morning, each utility estimates how much power from each source to expect the following day. This requires the day-ahead forecast that is accomplished by blending the NWP models via DICast®. Then in real-time, actual allocations must be adjusted and the Nowcast system is helpful for that process. Finally, the probabilistic forecasts are used by some utilities in planning their reserve allocations. Although most of the utility partners do not yet have a sufficient capacity of solar power for the forecasts to make a large difference in their operations, given the rapid growth of solar deployment in the U.S., they anticipate a time when such forecasts will be critical to their operations as wind power forecasts are already for many utilities.

V. Conclusion and Challenges

Renewable energy is becoming a higher percentage of the energy capacity as it becomes more prevalent and cost effective. Thus, it is important to forecast its expected value, variability, and uncertainty. To make such a forecast requires observations from that location, specialized models tuned to the location, and blending NWP data from multiple sources. Observations are additionally critical for building computational intelligence models that optimize the forecast. Therefore it is necessary to handle data that is large in volume, of high variety, using it in real-time at high velocity, and may be of questionable veracity. The complexity of blending this information in time to provide a forecast that is useful to the end user is a complex Big Data problem. Here we have described a solar power forecasting system, Sun4Cast™, developed by NCAR and collaborators that meets these needs. Numerous challenges must be overcome in making an accurate forecast.

Models are always being improved and resolution is growing as computer power increases. We sometimes hit limitations to our model resolution, however. As NWP models are deployed at higher resolution, we resolve processes that we previously parameterized. We are also modeling scales of the atmosphere beyond those for which the model was originally constructed, and thus, the important physics to resolve may change. For instance, although the NWP models were constructed to resolve large-scale flow over large domains, for much finer scales large-eddy simulations (LES) may be more appropriate. A "terra incognita" exists between about 1000 m and 100 m where the turbulence characteristics change, so that the type of model that is appropriate to use must be carefully considered [28].

As discussed above, we note that meteorologists often run multiple realizations of the NWP models as a way to quantify the uncertainty in the flow due to its chaotic nature. With bigger and faster computers available, modelers tend to run more ensemble members to better fill out the PDF of the forecast in an attempt to improve their predictions. Is that approach the best, or will statistical learning methods prove better at statistically filling out the PDF? The limiting case is demonstrated by the analog ensemble method, which uses a single high-resolution simulation to create the ensemble. This AnEn technique has shown promise for improving the deterministic forecast while quantifying its uncertainty.

More complex data mining techniques are being deployed to blend the models and the observations. For instance, deep machine learning is becoming more widely used in many Big Data problems. But those methods require yet more data. Again, we will observe how the balance between data volume and its smart usage develops over time.

Machine learning techniques depend critically on archived data for training. Thus, it becomes important to select and archive a subset of essential variables at appropriate frequency for future use. Similarly, the analog ensemble and other statistical learning methods, such as StatCast, require this type of historical data as well as coincident observations.

Because accurate renewable power forecasting so critically depends on the amount and quality of historical, real-time, and model data, it is important to consider archival at an early stage in project planning as well as for future projects. In addition to power production, the data regularly collected at solar power plants include atmospheric variables: e.g., GHI or DNI, temperature, wind speed and direction, as well as pressure and humidity. To effectively use these data, dataset formats should be standardized. In particular, comprehensive metadata is essential, including information about the instruments used to collect the data, accurate location information, the time of collection, and the instrument maintenance record. The location should be specified in latitude and longitude based on the World Geodetic System (WGS) standard from 1984, revised in 2004, and include the height above the surface. Our experience indicates that frequent confusion exists concerning local time zones and seasonal time changes due to daylight saving time, which can result in time lost by scientists and engineers to determine the correct time for the training data. To avoid this confusion, the time when the data were collected should be reported consistently in internationally accepted Coordinated Universal Time (UTC).

Our experience also indicates that dealing with arcane data formats wastes the time of the system engineers. Thus, we recommend that the data be organized and stored according to one of the established portable data formats for ASCII data; for example Met_Point, little_r formats, or self-describing binary formats such as GRIdded Binary or General Regularly-distributed Information in Binary (GRIB) format, Common Data Format (CDF, http://cdf.gsfc.nasa.gov/), Network Common Data Format (NetCDF, http://www.unidata.ucar.edu/software/netcdf/), or Hierarchical Data Format (HDF, http://www.hdfgroup.org/). Exploiting these well-defined, documented, and widely used data formats significantly enhances datasets' utility and simplifies their processing and quality control.

Finally, as such applications move toward cloud computing frameworks, additional complexities arise. As we deploy our models on a larger variety of architectures, the issues of disparate data arriving at different times and requiring blending to provide real-time forecasts will become more complicated.

Although these issues are challenging, the prospects for enhanced application are promising. As more wind and solar energy are brought into the energy grid, the need for accurate forecasts of the renewable energy variables grows. This demand for continually improving forecasts provides interesting research topics for atmospheric scientists and software engineers. Thus, solutions will continue to arise to meet the challenges.

Acknowledgment

The authors thank the entire Sun4Cast team at NCAR and beyond. Special thanks go to Seth Linden, who collected information on data usage in the Sun4Cast system, Matt Rogers who estimated daily data usage of the CIRACast system, Julia Pearson who prepared Fig. 1, Tara Jensen for preparing Fig. 4, and Jared Lee for providing an internal review. This work was supported by the U.S. Department of Energy under award number DE-EE0006016.

References

[1] G. Giebel, and G. Kariniotakis, "Best Practice in Short-term Forecasting - A Users Guide. European Wind Energy Conference and Exhibition," Milan (IT), 7-10 May 2007. Available: https://hal.archives-ouvertes.fr/hal-00525931/document.

[2] K.D. Orwig, M. Ahlstrom, V. Banunarayanan, J. Sharp, J.M. Wilczak, J. Freedman, S.E. Haupt, J. Cline, O. Bartholomie, H. Hamman, B.-M. Hodge, C. Finley, D. Nakafuji, J. Peterson, D. Maggio, and M. Marquis, "Recent Trends in Variable Generation Forecasting and Its Value to the Power System," *IEEE Transactions on Renewable Energy*, vol. 6, no. 3, pp. 924-933, 2014.

[3] W.P. Mahoney, K. Parks, G. Wiener, Y. Liu, B. Myers, J. Sun, L. Delle Monache, D. Johnson, T. Hopson, and S.E. Haupt, "A Wind Power Forecasting System to Optimize Grid Integration," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 4, pp. 670-682, 2012.

[4] M Ahlstrom, D. Bartlett, C. Collier, J. Duchesne, D. Edelson, A. Gesino, M. Keyser, D. Maggio, M. Milligan, C. Mohrlen, J. O'Sullivan, J. Sharp, P. Storck, and M. Rodriguez, "Knowledge is power: Efficiently integrating wind energy and wind forecasts." *IEEE Power and Energy Magazine*, vol. 11, no.6, pp. 45–52, 2013.

[5] C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann. "Wind Power Forecasting: State-of-the-Art 2009," Argonne National Laboratory, Argonne, IL, ANL/DIS-10-1, Nov. 2009.

[6] A. Tuohy, J. Zack, S. Haupt, J. Sharp, M. Ahlstrom, S. Dise, E. Grimit, C. Mohlren, M. Lange, M. Garcia Casado, J. Black, M. Marquis, C. Collier, "Solar Forecasting – Method, Challenges, and Performance, " *IEEE Power and Energy Magazine*, vol. 13, no. 6, pp. 50-59, 2015.
[7] J. Wilczak, and 17 coauthors, "The Wind Forecast Improvement Project (WFIP): A Public-Private Partnership addressing wind energy forecast needs," *Bulletin of the American Meteorological Society vol.* ,96, no. 10, pp. 1699-1718, 2015.
[8] S.E.Haupt and B. Kosovic, "Big Data and Machine Learning for Applied Weather Forecasts: Forecasting Solar Power for Utility Operations," IEEE Symposium Series on Computational Intelligence, Capetown, South Africa, Dec. 9, 2015.
[9] L.F. Richardson, "Weather Prediction by Numerical Process," Cambridge University Press, Cambridge, 1922.
[10] J.G. Charney, R. Fjortoft, and J. von Neumann, "Numerical integration of the Barotropic Vorticity Equation," *Tellus*, vol. 2, pp. 237-254, 1950.
[11] R. Hagedorn, R. Buizza, T.M. Hamill, M. Leutbecher, and T.N. Palmer, "Comparing TIGGE multipmodel foreasts with reforecast-calibrated ECMWF ensemble forecasts," *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 668, pp. 1814-1827, Feb. 2012.
[12] B. Kosovic, S.E. Haupt, G. Wiener, L. Delle Monache, Y. Liu, M. Politovich, J. Sun, J. Williams, D. Adriaansen, S. Alessandrini, S. Dettling, S. Linden. "Comprehensive Forecasting System for Variable Renewable Energy," International Conference on Energy and Meteorology, Boulder, CO, June 24, 2015.
[13] Z. Peng, D. Yu, D. Huang, J. Heiser, S. Yoo, and P. Kalb, "3D cloud detecting and tracking system for solar forecast using multiple sky imagers," *Solar Energy*, vol. 118, pp. 496-519, 2015.
[14] T.C. McCandless, G.S. Young, S.E. Haupt, and L.M Hinkelman, "Regime-Dependent Short-Range Solar Irradiance Forecasting," *Journal of Applied Meteorology and Climatology*, to be published.
[15] T.C. McCandless, S.E. Haupt, and G.S. Young, "A Regime-Dependent Artificial Neural Network Technique for Short-Range Solar Irradiance Forecasting," *Applied Energy*, vol. 89, pp. 351-359, 2016.
[16] T.C. McCandless, S.E. Haupt, and G.S. Young, "A Model Tree Approach to Forecasting Solar Irradiance Variability," *Solar Energy*, vol. 120, pp. 514-524, 2015.
[17] M. Rogers, S. D. Miller, J. M. Haynes, A. Heidinger, S. E. Haupt, and M. Sengupta, "Improvements in satellite-derived short-term insolation forecasting: Statistical comparisons, challenges for advection-based forecasts, and new techniques." *6th Conf. on Weather, Climate, and the New Energy Economy at the 95th AMS Annual Meeting*, Phoenix, AZ, 6 Jan 2015.
[18] T. Auligné, T., "Multivariate Minimum Residual Method for Cloud Retrieval. Part I: Theoretical Aspects and Simulated Observations Experiments," *Mon. Wea. Rev*., vol. 142, pp. 4383-4398, 2014.
[19] T. Auligné, T., "Multivariate Minimum Residual Method for Cloud Retrieval. Part II: Real Observations Experiments," *Mon. Wea. Rev*., vol. 142, pp. 4399-4415, 2014.
[20] G. Descombes, T. Auligné, H.-C. Lin, D. Xu, C. Schwartz, and F. Vandenberghe. "Multi-sensor Advection Diffusion nowCast (MADCast) for cloud analysis and short-term prediction," *NCAR Tech. Note*, Boulder, CO, NCAR/TN-509+STR, 21 pp, 2014.
[21] P. Jimenez-Munoz, S. Alessandrini, S.E. Haupt, A. Deng, B. Kosovic, J.A. Lee, and L. Delle Monache. "Role of Unresolved Clouds on Short-Range Global Horizontal Irradiance Predictability," *Journal of Applied Meteorology and Climatology*, to be published.
[22] P. Jimenez-Munoz, J.P. Hacker, J. Dudhia, S.E. Haupt, J.A. Ruiz-Arias, C.A. Gueymard, G. Thompson, T. Eidhammer, and A.J. Deng, "WRF-Solar: An Augmented NWP Model for Solar Power Prediction," *Bull. Amer. Met. Soc*., to be published.
[23] W. Myers, G. Wiener, S. Linden, and S. E. Haupt, "A consensus forecasting approach for improved turbine hub height wind speed predictions,"in *Proc. WindPower 2011*, Anaheim, CA, May 24, 2011.
[24] H. R. Glahn and D. A. Lowry, " The use of model output statistics (MOS) in objective weather forecasting," *J. Appl. Meteor*., vol. 11, pp. 1203–1211, 1972.
[25] S.E Haupt,, B. Kosovic, T. Jensen, J. Lee, P. Jimenez, J. Lazo, J. Cowie, T. McCandless, J. Pearson, G. Weiner, S. Alessandrini, L. Delle Monache, D. Yu, Z. Peng, D. Huang, J. Heiser, S. Yoo, P. Kalb, S. Miller, M. Rogers, and L. Hinkleman, "The SunCast Solar Power Forecasting System: The Results of the Public-Private-Academic Partnership to Advance Solar Power Forecasting," NCAR Technical Report, Boulder, CO, TN-526+STR, 307 pp, 2016.
[26] L. Delle Monache, T. Eckel, D. Rife, B. Nagarajan, K. Searight, "Probabilistic weather predictions with an analog ensemble," *Mon. Wea. Rev*., vol. **141**, pp. 3489-3516, 2013.
[27] S. Alessandrini, L. Delle Monache, S. Sperati, and J. Nissen, "Probabilistic power prediction with an analog ensemble," Submitted to *Renewable Energy*, 2016.
[28] J. C. Wyngaard, "Toward numerical modeling in the 'Terra Incognita'," *J Atmos Sci*, vol. 61, pp. 1816–1826, 20014.

Sue Ellen Haupt earned her Ph.D. in atmospheric science from the University of Michigan (1988), M.S. in mechanical engineering from Worcester Polytechnic Institute (1984), M.S. in engineering management from Western New England College (1981), B.S. in meteorology from Penn State (1978), and did a postdoctoral fellowship with the Advanced Study Program of the National Center for Atmospheric Research (NCAR). She is a Senior Scientist and Director of the Weather Systems and Assessment Program (WSAP) of the Research Applications Laboratory (RAL) of NCAR (Boulder, CO USA).

Dr. Haupt currently serves as a Director of the World Energy and Meteorology Council and is a Councilor of the American Meteorological Society (AMS). She previously served as Chair of the AMS Artificial Intelligence Committee and currently serves on the Energy Committee.

Branko Kosović received his Ph.D. degree in aerospace engineering from the University of Colorado (1996), M.S. degree in aerospace engineering from Penn State University (1991), and B.S. degree in mechanical engineering from the University of Rijeka, Croatia (1988). He is Program Manager for Renewable Energy for RAL and Science Deputy Director of WSAP (Boulder, CO USA).

Dr. Kosović chaired the Boundary Layers and Turbulence Committee of the American Meteorological Society (AMS) from 2006 to 2009. In 2010 he was a program chair for the 19th Symposium on Boundary Layers and Turbulence in Keystone, Colorado. He is currently serving as a member of the Renewable Energy Subcommittee of the AMS.