AN INTERNATIONAL SOLAR IRRADIANCE DATA INGEST SYSTEM FOR FORECASTING SOLAR POWER AND AGRICULTURAL CROP YIELDS

James Hall
JHTech
PO Box 877
Divide, CO 80814
Email: jameshall@jhtech.com

Jeffrey Hall
JHTech
PO Box 877
Divide, CO 80814

ABSTRACT

A data system has been developed for the real-time ingest of solar irradiance and climate data from thousands of ground-based measurement sites. Compared to space-based sensing of solar radiation, ground sensors can offer advantages in terms of accuracy and latency. The system is currently focused on the United States and the European Union, with data from other countries to be included in the future. It currently provides historical data for accurate solar resource assessment, real-time data for solar power forecast models, and data inputs for agricultural crop yield models. This ingest is unique in terms of the number of sites providing irradiance data, and the international scope of the project.

Outputs include continuous gridded and point global horizontal irradiance data for the day, continuous irradiance forecasts for the next four hours, and historical summaries of solar and climate data. The system archives the observations and includes routines to back-fill data that could not be obtained in real-time. Additionally the system monitors each sensor by analyzing the time history of observations, comparing measurements from nearby sites, and checking observations against theoretical solar envelopes.

Constructing a real-time ingest of ground-based data presented unique challenges such as asynchronous data availability, missing or delayed data, uncertainty about sensor maintenance and calibration, and the need to access many different networks with different data formats. This paper describes the methodology for addressing each of these problems and the current system performance.

1. INTRODUCTION

For the past decade solar radiation data from orbiting satellites and published historical datasets such as the US National Solar Radiation Database have been the mainstays in the design of solar power systems. Meanwhile many universities and government agencies around the world have built their own networks of ground-based solar and meteorological sensors for specific purposes such as agriculture, water management and environmental monitoring. By itself a single network may not provide much data; however, in aggregate they provide a significant resource.

Previous researchers have shown that ground-based measurements of daily irradiance are more accurate than space-based measurements up to a distance approximately 75 km from the site[1]. The challenge comes in aggregating a sufficient number of sites to completely and accurately represent the desired region. For the continental US (United States), this requires roughly 500 well-distributed measurement sites. Using the same criteria, the EU (European Union) would require about 250 sites to provide coverage superior to satellite-based measurements. In the US there are over 5500 publically accessible measurement sites along with thousands of sites in EU, and more are being added each year.

Depending on data logging and reporting methods, ground-based networks generally deliver data with significantly less latency than satellite observations. This reduced latency coupled with higher accuracy provides major advantages in developing systems for short-term solar radiation forecasts.

2. SYSTEM DESCRIPTION

A data system has been developed for real-time ingest of solar irradiance and climate data from thousands of US and international sites. This ingest is unique in terms of the number of sites providing irradiance data and the international scope. Figure 1 shows a graphical representation of the ingest system, and a brief description of the each component follows:

2.1 Data Sources

Currently the data ingest is focused on the US and the EU. Data from other countries will also be included in the future. Over 4500 sites in the US have been identified that report GHI (global horizontal irradiance) and other meteorological parameters in real-time. These sites are professionally operated and maintained by universities and government agencies and make their data publically available via the internet. Over 1000 similar sites have been identified in Western Europe. This data, however, must be gathered from over 100 disparate networks; each with unique formats, measurement units, sampling intervals and methods of data access.

Since each of these networks is independently operated for a specific local purpose, unexpected change is the norm. Typically each week one or more of the networks will modify their data access format. A major task in the operation of this system is on-going modification of the ingest software to adapt to these changes.

2.2 Ingest Controller

The goal is to acquire the data as near to real-time as possible. Each network has its own frequency of averaging and reporting the data, ranging from 5 minutes to one hour. Each network also has a different latency required to acquire the data from the sensor site, process the data and post it to the internet. If a scheduled data transmission is missed, the networks have different methods for acquiring and backfilling the observations. Based on these factors, one can estimate, but not know exactly, when the data from a given site should be available.

To make the ingest feasible, a scheduling controller is required that tells the system when to attempt to acquire data from a specific site. When the acquisition time for a site has arrived, the task is placed into a data acquisition queue. The processor scrolls repeatedly through this queue, attempting to acquire the desired data as soon as it becomes available. Once the data from a site is obtained, the actual acquisition time is logged and the data file is placed into a temporary folder for processing. If the data from a given site cannot be obtained within a specified time, the task is deleted from the queue until the next scheduled acquisition time. This ingest scheme scales linearly with the number of sites and the tasks are easily divided between multiple processors as required.

Each evening, a software process analyzes the attempted and actual data acquisition from each site. Based on this analysis, the trigger times to begin searching for each site's data and the delay before giving up the search are adjusted. This adaptive control process is crucial if the data ingest is to be run on a reasonable number of processors and within a reasonable internet bandwidth. It is also a key factor in respecting the providers of the data by not bogging down data access for other users.

2.3 Data Format Conversion

Each meteorological network, and often individual sites within the network, has its own data structure and measurement units. Processing software examines the header information for each ingested data file and compares it to the expected format for the site. If the data header matches, the data is converted to a common data format using a software routine specifically written for that network or site. If the header does not match, the file is transferred to an exception queue for manual examination. As previously mentioned, a major operational task is adapting the data conversion software to handle these recurring changes in the input formats.

A second, key function of the data converter is to place an accurate time stamp on each observation. To this end, the converter must keep track of the sampling frequency for each site and the ending time of the observation. Since the data is ingested from many different time zones, it is converted from the local time to UTC time (coordinated universal time), including adjustments for daylight savings.

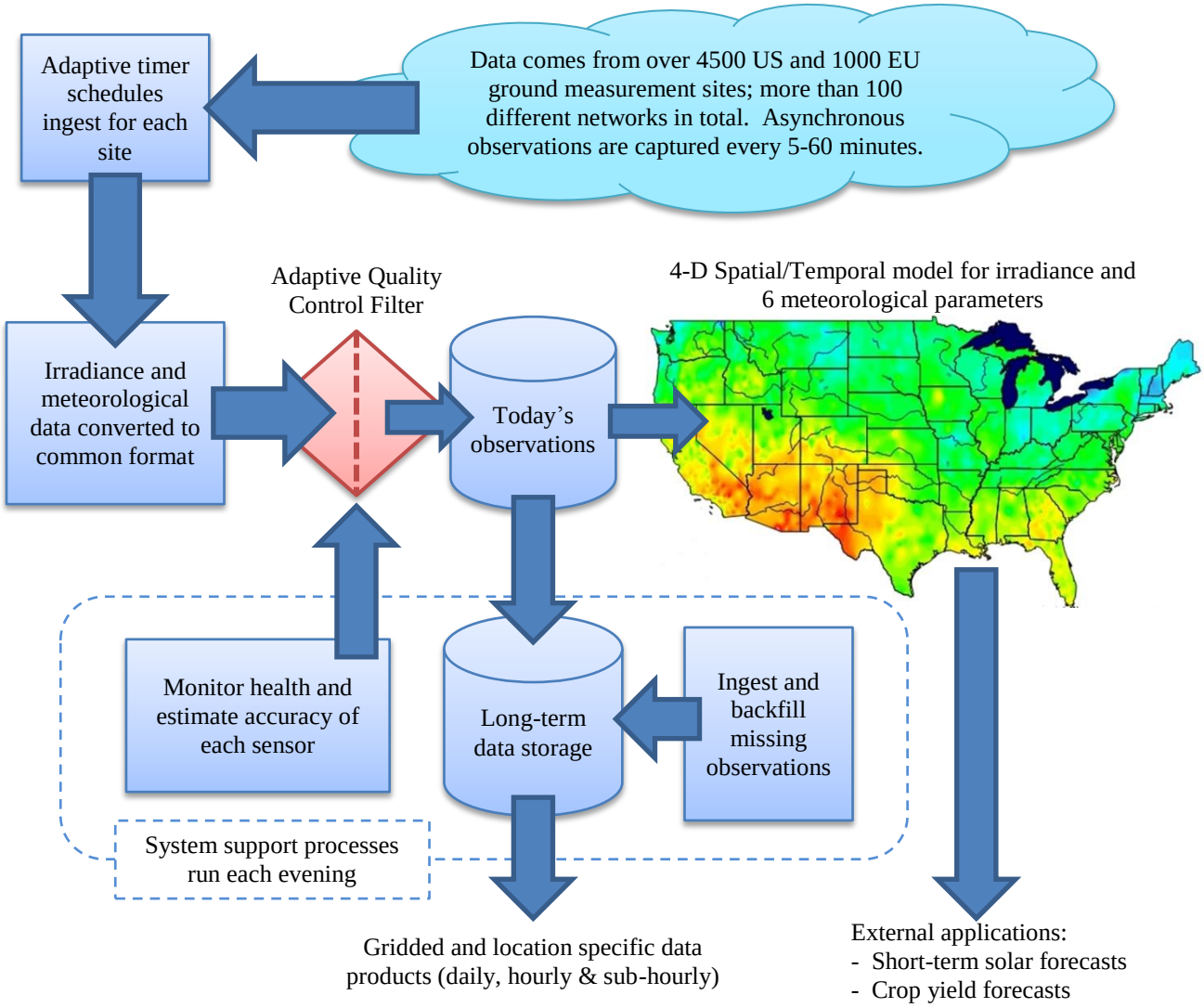

Fig. 1: Real-Time Data Ingest System for Solar Irradiance and Climate Data

2.4 Quality Control Filter

With data from many thousands of sensors being ingested each hour, it is likely that some of the sensors are not functioning correctly at a given time. The purpose of the quality control filter is to monitor the data flow from each site and pass only data from sensors that are deemed to be well calibrated and operating correctly.

During daylight hours, the filter is simply a set of flags, one for each sensor, which indicates if the current data from that sensor should pass to the modeling process or be rejected.

The flags are set every evening by analyzing the recent data streams from the sensors. For each parameter measured, a sliding history of recent data is used to judge its reliability. This judgment includes three elements: (1) Based on date and time of day, are the observations within reasonable statistical range of the long-term historic values? (2) Are the observations reasonably close to the readings from nearby sites at similar elevations? (3) At times identified as clear sky events, are GHI readings reasonably close to the expected clear sky radiation?

Sensors that are blocked are re-examined each evening to determine if the problem has been corrected and the data block can be removed.

2.5 Database of Current Observations

All observations that pass the quality control filter are stored sequentially as received into a database for the current day. This database contains seven separate tables for GHI, temperature, relative humidity, wind speed, wind direction, precipitation and soil temperature. The data structure for each table is straightforward: latitude, longitude, observation time (UTC), and the observed value.

2.6 Spatial-Temporal Modeling

The spatial-temporal models are 4-dimensional models created each day for each of the seven parameters (tables) in the current observation database. The four dimensions are latitude, longitude, time and parameter value.

To understand what this might look like, one could visualize a movie of 3-D surface plots showing solar radiation (GHI) at given time slices throughout the day. Frames of the movie before the current time would be generated by fitting a surface to the observations for that day. Frames of the movie into the future are time-based extrapolations of the past surfaces. Of course, the 4-D surface is actually continuous in time. This provides a simple solution to handling asynchronous data from a variety of physical locations. Any missing observations are smoothed over by the modeling surface and any delayed observations are inserted at the proper point in time.

Perhaps the most compelling reason for this technique is that the spatial-temporal surface inherently contains short-term forecasts several hours into the future. These short-term forecasts are determined by the current state as well as the historical dynamics of the spatial-temporal surface.

3.0 Long-Term Data Storage and Data Products

Each evening the database of current observations is transferred to long-term data storage. Missing observations from each site are flagged, and a software routine attempts to backfill the missing data for several successive days. This long-term database becomes the source of location-specific products, such as historical ground-based GHI observations for a planned solar project.

The spatial-temporal grid for the day is also saved to another database. This becomes the source for gridded data products, such as sub-hourly GHI and climate data for any desired latitude and longitude.

4.0 SHORT-TERM SOLAR FORECASTS

One application of the real-time ingest system is short-term forecasting of solar power. By short-term we mean forecasts one to four hours into the future. These forecasts are becoming increasingly important to power grid operators who must incorporate significant quantities of PV (photo-voltaic) power into the grid. They are also becoming important to owners of large-scale PV plants that must contract in advance to deliver a specific amount of PV power.

The real-time data ingest described above provides a fundamental advantage to any forecasting system. Since all measurements are ground-based and carefully quality controlled, they provide a more accurate starting point with less latency than with any other current technology.

Figure 2 shows a graphical representation of the production short-term solar forecasting system currently under development. This system is based on a prototype for short-term forecasts of GHI for the Los Angeles Basin[2].

The first forecasting module comes from the spatial-temporal model of solar radiation built into the data ingest system. As discussed above, this model extrapolates GHI forward in time and inherently combines several traditional methods of solar forecasting such as persistence, cloud vector analysis and other dynamic effects.

The second forecasting module is based on recognition of daily solar radiation patterns. The long-term database provides historical daily solar radiation patterns for the season and region of interest. These historical patterns are pre-generated from a sliding window of observations and are stored in a pattern file. As the day progresses, the system finds the closest pattern match and forecasts solar radiation based on a known historical pattern.

The third forecasting module is based on meteorological parameters. Again the long-term database provides the meteorological observations for the region and the resulting solar radiation. Similar to traditional weather forecasting, this process attempts to find the relationships between current weather observations and future solar radiation.

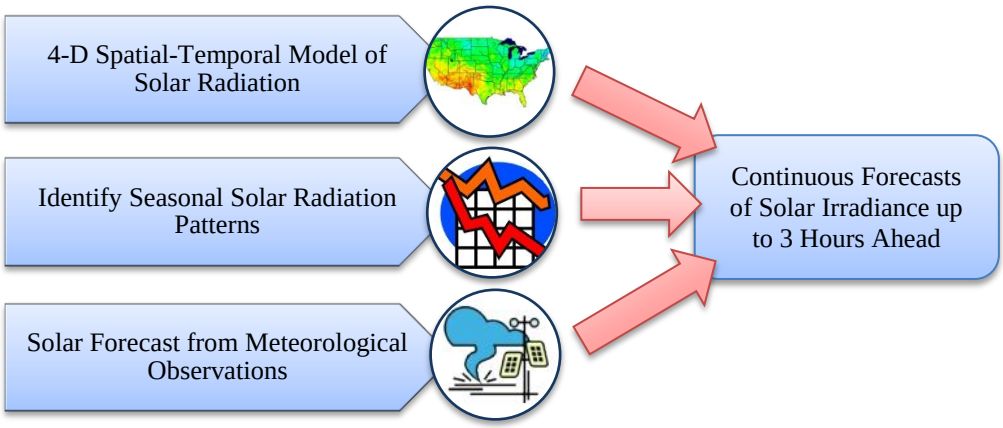

Fig. 2: Three Component Ensemble for Short-Term Forecasts of Solar Irradiance

The final short-term forecasts are a weighted combination of the forecasts from the individual modules. This ensemble approach provides robustness to the process and makes it simple to incorporate additional forecasting techniques into the system. The final weighting factors are determined from experience and will vary by season and region.

5.0 CROP YIELD FORECASTS

Over the past decade, researchers have demonstrated that when the micro-climate in which a plant grows is well known, final crop yields can be predicted quite accurately. The required data includes knowledge about genetic traits and soil type, as well as accurate daily (or preferably hourly) measurements of soil temperature, precipitation, air temperature, relative humidity, wind speed and GHI. With the exception of GHI and genetic traits, all of this data is relatively easy to obtain.

In addition to ingesting and compiling solar radiation data, we have compiled an extensive database of crop genetic traits and the corresponding yields. These two datasets will enable the daily forecasting of the final corn, soybean and wheat yields for growers at the local level and for US commodity markets at the national level.

6.0 OPERATIONAL STATUS AND FUTURE PLANS

The basic system described in Figure 1 is currently operational and ingesting real-time data from over 2500 US sites and 200 EU sites. We anticipate that by fall 2012 we will be ingesting our initial target of 4500 US and 1000 EU sites. The spatial-temporal model is operational and producing short-term forecasts of solar radiation. The sensor monitoring component of the quality control filter is not yet operational, be we anticipate that it too will be completed by fall 2012.

A significant portion of the development effort to date has focused on obtaining complete coverage of California, for which 550 observation sites are currently being ingested. Spatial-temporal forecasts for California are now available and two additional forecasting components are partially complete: the pattern recognition component and the meteorological component. We anticipate that full production forecasts of solar radiation for California will be available summer of 2012.

7.0 CONCLUSIONS

This paper describes a new ingest system focused on ingesting and processing solar irradiance data for the US and the EU. All data are collected in real-time from a dense network of ground-based sensors. Each observation is carefully controlled by multiple methods. We believe that this system will provide the most accurate and timely view of solar irradiance that is currently available.

8.0 REFERENCES

(1) Perez, R.,Kmiecik, Zelenka, A., "Determination of the Effective Accuracy of Satellite-Derived Global, Direct and Diffuse Irradiance in the Central United States", [Online]: http://www.asrc.cestm.albany.edu/perez/directory/ResourceAssessment.html.

(2) Hall, J., Hall J., "Forecasting Solar Radiation for the Los Angeles Basin – Phase II report", Presented in SOLAR 2011, The National Conference of the American Solar Energy Society (ASES). Raleigh, NC, May 17-21, 2011.