Pacific Northwest
NATIONAL LABORATORY

PNNL-29375

Data Assessment and Assimilation for Atmospheric Radiation Measurement Data Using Dynamic Bayesian Networks

November 2019

George Chin Jr
Juan M Brandi
William I Gustafson
Katie Porterfield
Laura D Riihimaki

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-29375

Data Assessment and Assimilation for Atmospheric Radiation Measurement Data Using Dynamic Bayesian Networks

November 2019

George Chin Jr
Juan M Brandi
William I Gustafson
Katie Porterfield
Laura D Riihimaki

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Data Assessment and Assimilation for Atmospheric Radiation Measurement Data Using Dynamic Bayesian Networks

George Chin Jr., Juan M Brandi, William I. Gustafson, Katie Porterfield, Laura D Riihimaki

◆ A data-driven machine learning approach to scientific sensor data assessment may automatically learn directly from the structural and temporal patterns in abundant observations. The project has developed data assessment and assimilation models to screen datasets using dynamic Bayesian network (DBN) and deep learning (DL) methods, which rely on correlations between variables, across time, and between spatial locations, to detect poor or invalid data with a high degree of confidence. ◆

The increasingly data-driven nature of scientific disciplines makes data quality assessment a first-order problem. For instance, the DOE Atmospheric Radiation Monitoring (ARM) Facility operates one of the largest climate research programs dedicated to the collection of long-term continuous measurements of cloud properties and other key components of the earth's climate system. Given the critical role ARM data plays in the analysis of the atmosphere and in the enhancement and evaluation of climate models, the production and distribution of high-quality data is one of ARM's primary mission objectives.

A common, reoccurring task when interfacing observations with scientific models is verifying the integrity of datasets. This is both time-consuming and prone to subjective choices. However, it is also common in many arenas to have well-sampled conditions with interrelated measurements that can inform the validity of each other. In environments where abundant observations are available, such as ARM, a data-driven machine learning approach to scientific sensor data assessment should prove to be effective as one can automatically learn directly from the structural and temporal patterns in the abundant observations.

On the project, a sensor-based DBN was initially developed as a data assessment, cleansing, and assimilation model. The sensor-based DBN utilized the following input sources:

- Date and time
- Temperature
- Wind speed and direction
- Relative humidity
- Height and elevation
- Barometric pressure
- Precipitation
- Vapor pressure
- Mean radiation

The model outputs were to predict the following sky brightness temperatures:

- Tbsky23
- Tbsky30
- Tbsky31
- Tbsky89
- Tbsky23_mwr3c

These measurements help to determine the liquid water and vapor contents suspended in the atmosphere.

In the DBN, the variables' conditional probability tables (CPTs) are determined by the number of parents and all their possible combinations of states. Because of this, when there is a large number of variables and/or a large number of conditional dependencies, CPTs can grow prohibitively large and computationally expensive. This was our case, where for example, the mean radiation measurement alone required measurements at about 2,000 different frequencies.

For DBN parameter learning, ARM data from the year 2016 was fitted using the *maximum likelihood estimate* method. An 85/15 split of the data was used for training and validation. In validation, we compared predicted versus actual sky brightness temperatures for two of the variables over a month of data. Validation loss was calculated using mean squared error (MSE) over the validation set of the five sky brightness temperature variables. The evaluation set achieved a mean square error of 63.4, which represents a moderately inaccurate result.

In an effort to improve accuracy and computational performance, a parallel effort was pursued to implement a deep neural network (DNN) model as an alternative to the DBN approach. A DNN is a neural network that consists of one input layer, two or more hidden layers, and one output layer. DNNs are a class of deep learning methods. Each DNN's hidden layer captures and transforms more abstract features from the data.

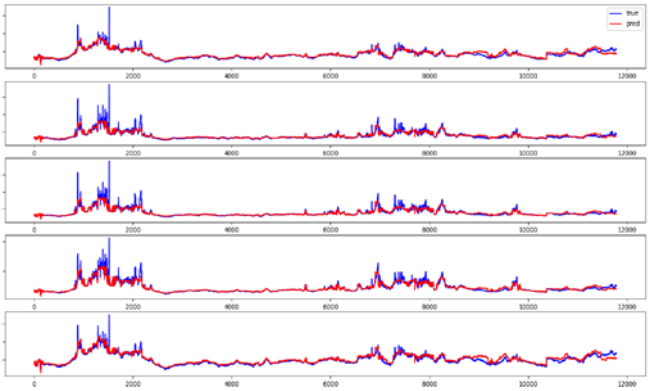
We took advantage of the flattened data processed for the DBN and created a fully-connected DNN composed of three hidden layers with 4,330, 2,165, and 433 units respectively. We used the python

wrapper module Keras with Google's TensorFlow machine learning library as the backend to create and train the DNN. Using TensorFlow has the computational advantage of using multiprocessing interfacing with the computer's graphics process units (GPUs). This translates into faster training and evaluation times. An Adam optimizer was used to calculate the gradients and the rectified linear unit (ReLU) was used as the activation function. Validation loss was calculated using MSE over the validation set.

In general, the DNN was more accurate with the evaluation set with a MSE of 4.7 compared to 63.4 achieved by the DBN. However, the DNN is only processing cross-sectional data and not capturing temporal relationships. In order to improve accuracy further and look at temporal relationships, we created a model based on another class of DL methods known as long short-term memory (LSTM).

LSTMs are useful for processing time-series data where dependencies exist over time. They can learn the arbitrarily long-time dependencies on a data sequence. LSTMs usually consist of an input gate, a forget gate, and an output gate. LSTMs "remember" their state from the previous time step and use it along with the current time step as input.

Our latest DL model has the LSTM layer with 236 units and ReLU activations. The LSTM model was more accurate than both the DNN and DBN models with a MSE of 3.7.



Pointwise comparison of the five sky brightness temperature variables produced by the LSTM model. Red plotline shows estimated LSTM values over time, while blue plotline shows observed values. Mean squared error (MSE) for LSTM model is 3.7.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov