Data Network Weather Service Reporting
DE-FG02-09ER25909/DE-SC0001661

- Final Report -

Date: August 30, 2012

Award No.: DE-SC0001661

Award Recipient: Michael Frey, Bucknell University

Report Period: September 1, 2009 – August 31, 2012

To: Thomas Ndousse-Fetter, Program Manager, DOE, Office of ASCR

Prepared by: Michael Frey
Department of Mathematics
Bucknell University
Lewisburg, PA 17837
mfrey@bucknell.edu
570.577.1598

Technical Summary: This final report covers the full period of a three-year effort to create statistical components for the e-Weather Center. The major accomplishments of this effort are:

- Adaptation of Gaussian process regression to network performance characteristic (NPC) forecasting
- Development of a new and novel NPC forecasting tool called flow field forecasting
- Flow field forecasting coded in the Python scripting language
- Integration of flow field forecasting into e-Weather Center
- Promotion of flow field forecasting's availability to diverse communities.

Technical Details: Year 1 (9/2009 – 8/2010) of this report's three-year period proceeded in three stages: 1) setting desiderata for e-Weather Center forecasting, 2) comparing and testing candidate statistical forecasting procedures against these desiderata, and 3) coding the selected forecasting procedure for integration into the e-Weather Center.

Year 1 - Stage 1: The goals for e-Weather Center were used to derive key desiderata for statistical forecasting of network performance characteristics (NPCs):

- Forecasts should be accompanied by a margin of error.
- Forecasting should require no supervision by an operator.
- Forecasts should involve no open-ended numerical searches.
- Forecasting must accommodate missing data and data collected aperiodically.
- The forecasting procedure's computational complexity should grow only modestly with the amount of NPC data.

Year 1 - Stage 2: Two mature theoretical frameworks, neural network forecasting and Box-Jenkins ARIMA modeling, initially seemed most promising for NPC forecasting, but each was subsequently rejected for failure to meet significant e-Weather Center goals. Gaussian process regression (GPR), which researchers have only in recent years developed sufficiently for forecasting, was considered at this stage to best meet e-Weather Center's needs.

Year 1 - Stage 3: A version of GPR adapted to e-Weather Center forecasting was designed and coded in C++ with MatLab modules. This software was delivered to the e-Weather Center integration team at FermiLab in July 2010. The FermiLab team for a variety of reasons did not undertake to integrate the forecasting software at this stage, and GPR was ultimately never integrated into e-Weather Center.

Year 2 (9/2010–8/2011) of the project proceeded along two parallel tracks: 1) integrate forecasting into e-Center and 2) develop anomaly detection tools to complement those under development by Prasad Calyam at OSU.

Year 2 - Track 1: The FermiLab e-Weather Center integration team agreed that forecasting software would be written in the open-source scripting language Python. Meanwhile, ongoing testing with GPR and efforts at its improvement led to the development of a completely new and novel forecasting paradigm, named flow field (FF) forecasting. FF forecasting uses penalized spline regression to build a skeleton of the structurally determined part of the NPC data. Then, using GPR, FF forecasting extends this skeleton (or flow field) into the future to make forecasts. Direct forecasting with GPR (as originally pursued in Year 1) is dogged by the technical challenge of modeling the NPC covariance. FF forecasting replaces this problem with the far easier task of modeling the covariance in the flow field. Equally well, FF forecasting recasts the difficult problem of *extrapolation* as a sequence of easier *interpolations*. Unlike GPR, FF forecasting makes only the very mild inductivism assumption that, under similar conditions in the future, NPC changes will tend to be similar to those seen in the past. FF forecasting readily meets all the e-Weather Center forecasting desiderata (listed above). An initial Python coding of FF forecasting was written, and testing and refinement continued through the summer of 2011. FF forecasting coded in Python was delivered to the e-Weather Center integration team in August 2011, and integration with e-Weather Center was completed soon after.

Year 2 - Track 2: An anomaly is an occurrence in an NPC that is strange, unusual, or unique; it is any discrepancy or deviation from an established pattern or trend. Anomaly detection is a foundation for visualizing and interpreting network dynamics. One approach to anomaly detection is to create distinct detectors—essentially control charts borrowed from the field of engineering statistics—for each of the many kinds of anomalies that might possibly present themselves in NPC data. This is the approach taken by Prasad Calyam at OSU, with some impressive results. A complementary approach is to seek a robust general detector that can recognize any change in the structure of the data. Prof. Kyle Caudle at the South Dakota School of Mines & Technology has been working with wavelet decompositions of data streams to detect anomalies. We collaborated to adapt and refine this approach to the problem of NPC anomaly detection. This collaboration proved useful, and some initial tests were performed.

Year 3 (9/2012–8/2012) of the project was committed to further development of FF forecasting and its promotion to various communities. Prasad Calyam's efforts at OSU toward anomaly detection proved effective, enough so that development of wavelet decompositions begun in Year 2 to detect anomalies was discontinued. This wavelet approach to anomaly detection remains for future study.

Year 3 – Continuing development of FF forecasting: FF forecasting is performing effectively within e-Weather Center, having been tuned specifically for this purpose. For subsequent applications in other arenas, a more capable, self-tuning version of FF forecasting would be welcome. To be self-tuning, FF forecasting needs capabilities to

- Automatically choose appropriate variables for the history space. This can probably be done via principle components.
- Estimate the correlation scale in the history space directly from the data. Simple method-of-analogy estimation appears to be effective for this problem.
- Update FF forecasts in a simple fashion as new data arrive without redoing all the calculations. It appears that FF forecasts can be updated by a new technique we call stitching.

The development of these self-tuning capabilities remains to be accomplished.

Year 3 – Promotion of FF forecasting: FF forecasting is applicable to forecasting problems well beyond the domain of NPC forecasting. Statistical forecasting is broadly recognized to be challenging, and a new, effective, broadly applicable tool for forecasting should be very welcome. To introduce diverse audiences to the potential of FF forecasting and to stimulate broader efforts at its development, presentations were made at the following meetings:

- Army Conference on Applied Statistics, Annapolis, MD, 19–21 Oct. 2011
- 10th International Conf. on Machine Learning, Honolulu, HI, 18–21 Dec. 2011
- International Symposium on Forecasting, Boston MA, 24–27 June 2012
- Joint Statistical Meetings, San Diego, CA, 28 July–2 Aug. 2012

A number of commercial enterprises—PayPal, Hershey Co. and the San Diego Padres—have expressed interest in FF forecasting, and potential collaborations are being explored.

Four papers have been published on FF forecasting:

- M. Frey and K. Caudle, "A New Premise for Forecasting," *Proceedings: 17th Annual Army Conference on Applied Statistics*, Annapolis, MD, 19–21 Oct. 2011.
- M. Frey and K. Caudle, "Introducing Flow Field Forecasting," *10th International Conf. on Machine Learning*, Honolulu, HI, 18–21 Dec. 2011.
- M. Frey, K. Caudle and Y. Liu, "Characteristic Length Estimation for Flow Field Forecasting," *Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, 24–27 June 2012.
- K. Caudle and M. Frey, "Continuous Updates of Penalized Spline Regression for Flow Field Forecasting," *Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, 24–27 June 2012.

Two more papers are being drafted, one for the *2012 Proceedings of the JSM* and one for *Statistical Analysis and Data Mining*.

Budget Summary: This project covered a three-year period (9/1/09–8/31/12). Award amounts budgeted and expended in each of the three years are shown in Table 1.

	Year 1 (9/1/09–8/31/10)	Year 2 (9/1/10–8/31/11)	Year 3 (9/1/11–8/31/12)	Total for 3-year period of award
Expended	$84,143	$62,935	$86,339	$233,417
Budgeted	$113,364	$118,384	$123,651	$355,399
Surplus	$29,221	$55,449	$37,312	$121,982

Table 1: Award budgeted and expended amounts (to the nearest dollar).

Table 1 shows that of the $355,399 budgeted for the three years of the award, $233,417 (66%) were expended, leaving a surplus $121,982 (34%) unexpended. Some unusual events account for the underutilization of funds. Announcement of the award decision was delayed three months past the 9/1/2009 start date, closing opportunities in Year 1 for hiring. The long delay involved in resolving with the e-Weather Center integration team to use Python to code the forecasting algorithms was also a factor. Additionally, Python is free open-source software and its unanticipated choice for this project saved funds. Finally, extensive visits to FermiLab to coordinate efforts with team members there turned out to be unneeded, and the planned funds for those trips were not used.