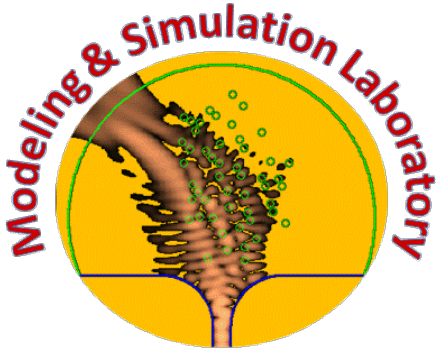
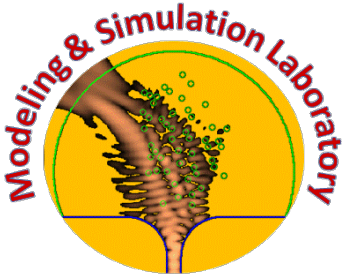


Introduction to A Space Weather Forecast Test-bed

Chunming Wang

Modeling & Simulation Laboratory
Department of Mathematics
University of Southern California

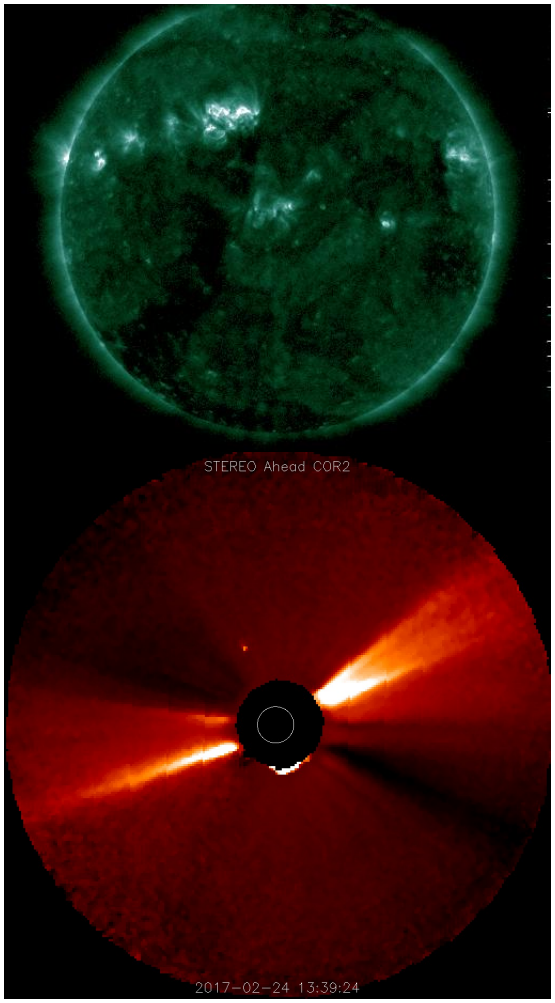


Development of Forecasting Strategies Is Driven by Physics and Empirical Evidence

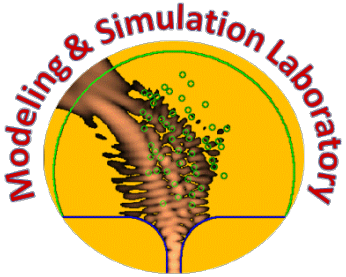
- Laws of physics provide links between solar and interplanetary magnetic field (IMF) events to thermosphere and ionosphere variations.
  - Available data are insufficient for models calibration.
  - Models are most reliable for quiet and regular conditions.
- Ample historical evidences exist to connect solar and IMF anomalies to ionosphere disturbances.
- Successful forecast must leverage all available information.
  - Select features of ionosphere most susceptible to solar and IMF.
  - Combine empirical data and models to forecast extreme events.

Extreme Events are Most Valuable and Challenging to Forecast

Systematic Mining of Historical Space Weather Data is Necessary

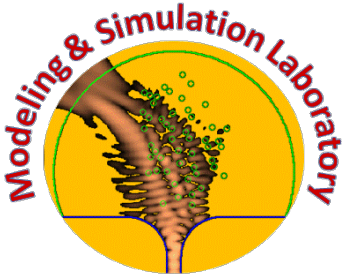


- Large volume of historical space weather data including solar imagery can reveal important clues through systematic analyses.
- Analyses focused on isolated events may overlook relevant features.
- Machine learning techniques for image recognition and features extract are very advanced.



A Test-bed is an Incubator for Space Weather Forecasting Strategies

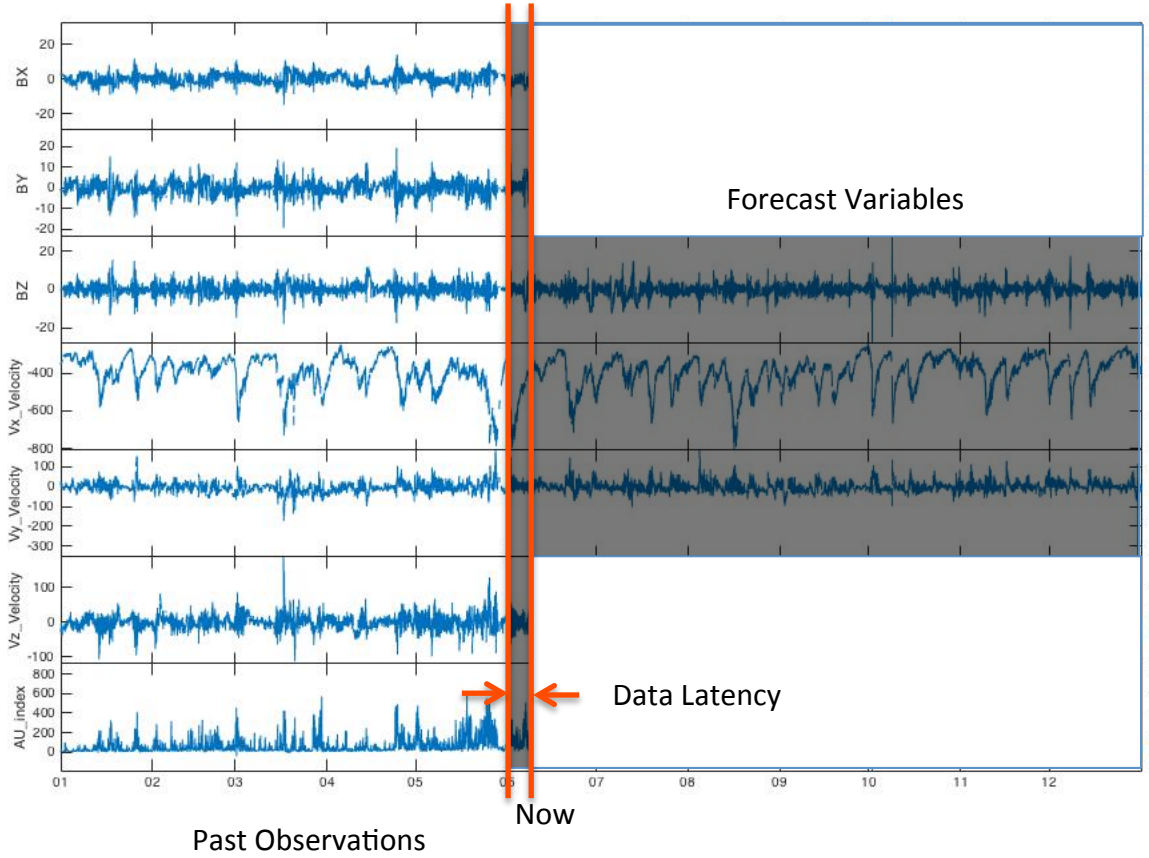
- A database of high quality historical space weather observations enables training and validation of forecast strategies.
- A collection of empirical and first principle physics models, as well as, data analysis tools facilitates development of new forecast approaches.
- A powerful computational and data storage platform provides capability to analyze large scale of historical data and leveraging previously computed intermediate results.



Outline

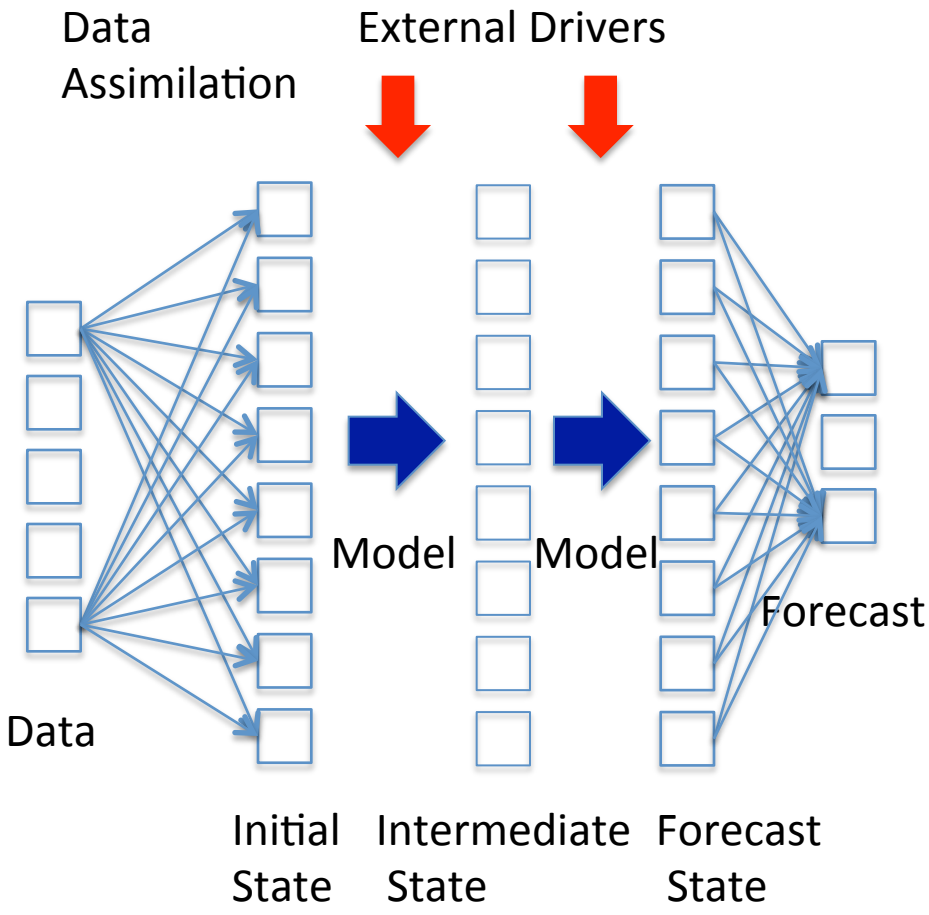
- Vision and Objective of the Test-bed
  - Paradigm and metrics of forecasting.
- Historical Database
  - Raw data and quality control of space data.
  - Ionosphere data and feature extraction.
- Regression Analyses
  - Forecasting strategies based on correlation and auto-correlations.
  - Training and validation of regression based forecast.
- Support and continued development of the test-bed
  - Potential data sources and other forecasting strategies.

Forecasting Relies on Connections Between Past and Future Events



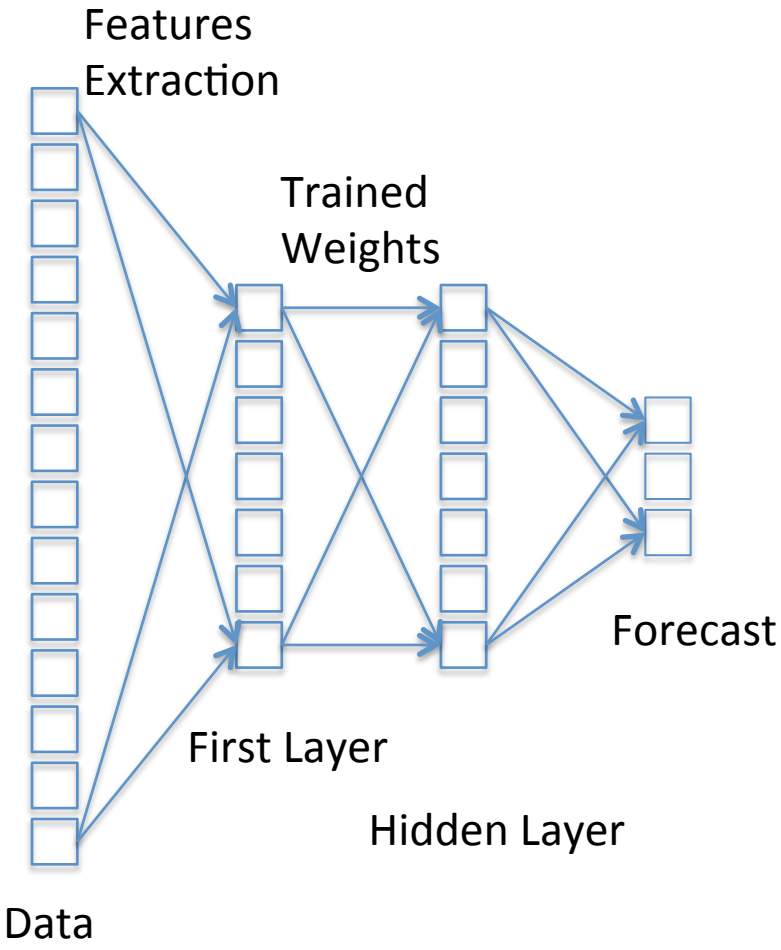
- Forecast variables are observable.
- Data used in producing a forecast may include many observations that are not necessarily important to forecast.
- Data latency in forecast is present.

Physical Laws Involve Internal State to Link Present to Future

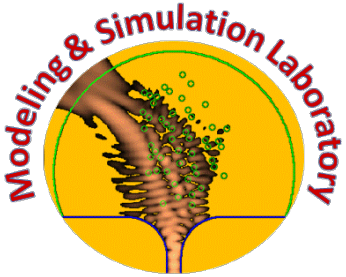


- Model based forecast system propagates internal state of the system.
- Data is used to determine initial state.
- Forecasts for external drivers are needed.
- Models indirectly link past data to forecast.

Most Data Mining Approaches Recognize the Need of Hidden State



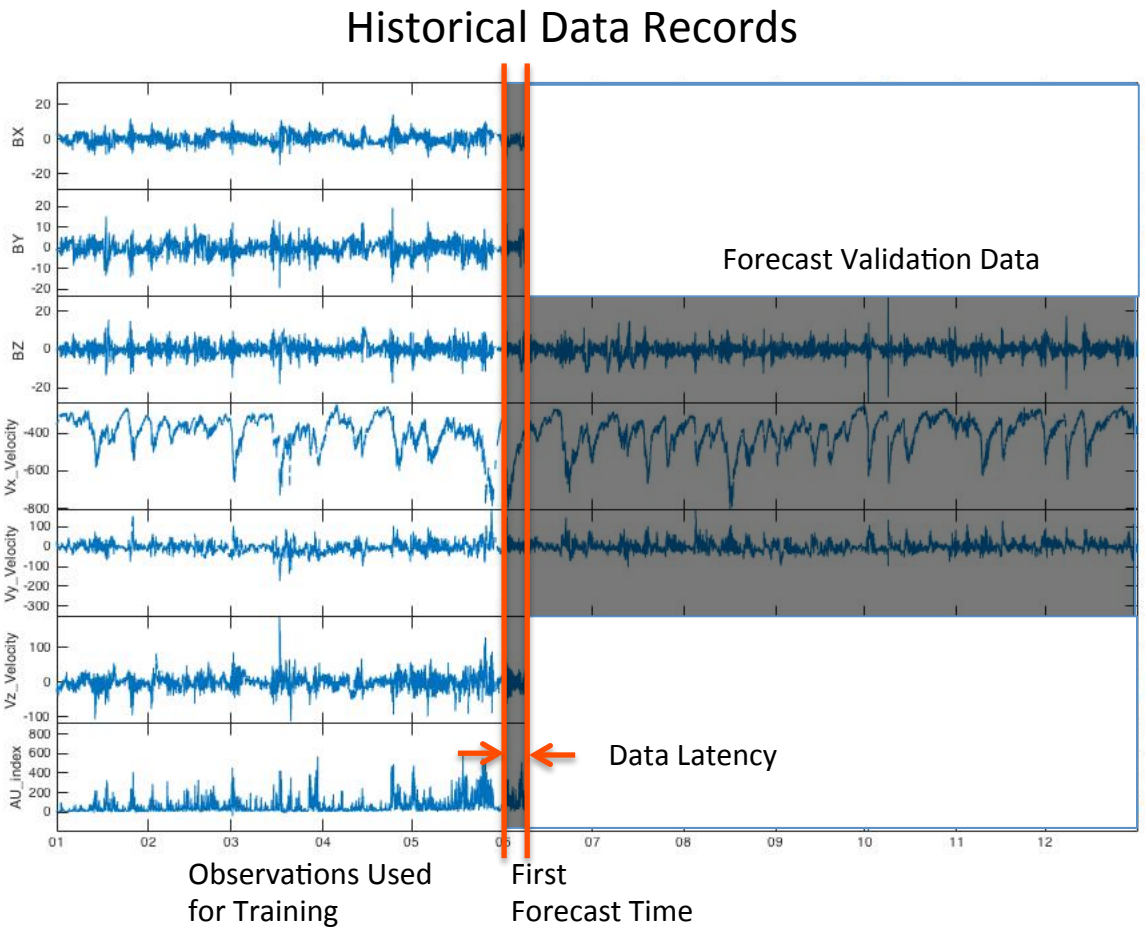
- In an empirical model, all relationships are derived from historical data.
- Physical insights help in feature extraction and determine the general structures of the model.
- Statistics and optimization are key in training the models.



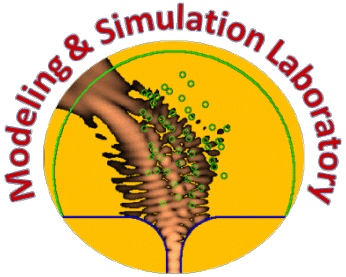
Combination of Physics Based and Empirical Models is Most Promising

- Space weather forecast is most likely to be successful when physics based models are complemented with data driven machine learning tools.
  - Physics based models can be used as a filter to extract most relevant features in data.
  - Physics model can also be used as constrain for model built by machine learning techniques.
  - Available observation data anchor the forecast system to the reality.
- Combined approach is at the cutting edge of research.

Forecast Test-bed Provides a Community Platform for Exploration

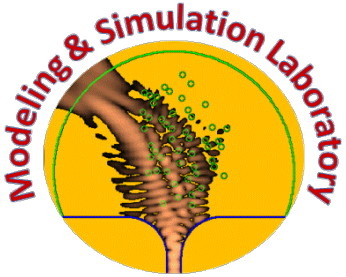


- Development of all forecast systems requires combinations of physical insights and experiences.
- Quality control and preparation are needed.
- Shared tools and experiences stimulate creativity.



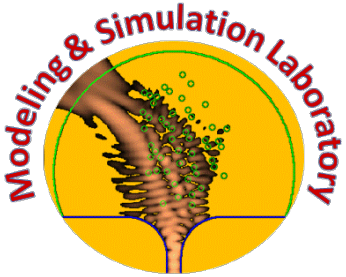
A Test-bed Can Help Forge Consensus on Values and Metrics of Forecast

- What to forecast can be just as important as how to forecast.
  - Characterization of thermo-ionosphere anomalies through outliers detection.
  - Definition of anomalies based on their impacts which must be consistently measurable.
- Defining a value for a forecast is useful to help focus efforts on high impact events.
- Objective metrics is critical for demonstrating progresses.



An Easily Accessible Test-bed Attracts Talents to Space Weather Forecast

- Data mining and machine learning are fields of research that have potential to impact many areas.
- Skills developed by graduate students in space weather data analyses and forecast can be useful for a variety of career paths.
- Interdisciplinary collaborations offer the best chance for break through.
  - Leveraging advances in data sciences benefits space weather forecasting.



Outline

- Vision and Objective of the Test-bed
  - Paradigm and metrics of forecasting.
- Historical Database
  - Raw data and quality control of space data.
  - Ionosphere data and feature extraction.
- Regression Analyses
  - Forecasting strategies based on correlation and auto-correlations.
  - Training and validation of regression based forecast.
- Support and continued development of the test-bed
  - Potential data sources and other forecasting strategies.

Data Sets in SWFT Are Collected from Different Sources

- The raw data included in SWFT are from 3 data providers
  - World Data Center for Geomagnetism, Kyoto, http://wdc.kugi.kyoto-u.ac.jp/dstdir/
  - NGDC/NOAA, ftp://ftp.ngdc.noaa.gov/STP/GEOMAGNETIC_DATA/INDICES/KP_AP
  - Goddard Space Flight Center, Space Physics Data Facility's Omniweb: http://omniweb.gsfc.nasa.gov/form/omni_min.html

Space Weather Data in SWFT Include Observations and Derived Indices

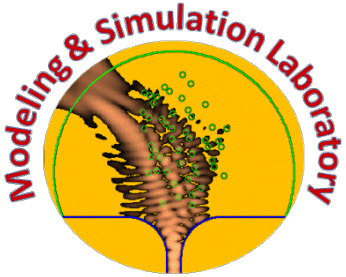
- Observations of near-Earth solar wind, magnetic fields and plasma are retrieved from multiple missions.
- Sun-spot, F10.7 and 3 hour geomagnetic indices AP, Kp are keys indices used by many models.
- Disturbance Storm-Time (Dst) Index derived from a network of near equatorial geomagnetic observatories is expected to be strongly correlated with thermo-ionosphere disturbances.

Common Temporal Grid of 3 Hour Resolution is Used for All Variables

- To facilitate analyses, all data are resampled to 3 hour time resolution.
  - High resolution solar wind and magnetic field data are represented by their summary statistics over 3 hour intervals such as median, min/max and total variation.
  - Daily data such as Sun-spot and F10.7 are constant throughout the day.
  - Hourly Dst is represented by its median, min/max and total variation over 3 hours intervals.

Global Ionosphere Conditions Are Represented by GIMs

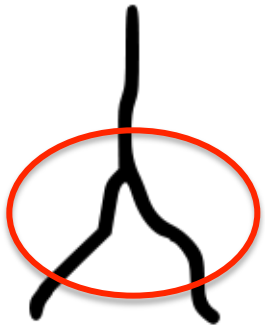
- At an initial stage the conditions of the ionosphere is represented by the Global Ionosphere Map (GIM) produced by JPL.
  - GIMs are maps of Vertical Total Electron Content (VTEC) converted from GPS ground receiver data.
  - A global map of 1 degree in latitude and longitude is produced every 15 minutes.
  - GIMs have been continuously generated for over 20 years.
- Additional global, regional or local ionosphere observation can be added to SWFT.

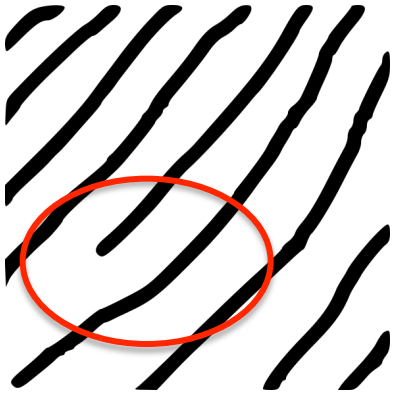


Only Key Features of GIM Are Interesting to Forecast

- Due to the sparsity of ground GPS receiver network used in the derivation of GIM, some artifacts of data extrapolation exist.
- As 2D VTEC map, a GIM cannot fully represent the state of ionosphere.
- Only large scale features in GIM can be expected to be correlated to solar, interplanetary magnetic field and geomagnetic field disturbances.
- Extraction of key features are crucial.

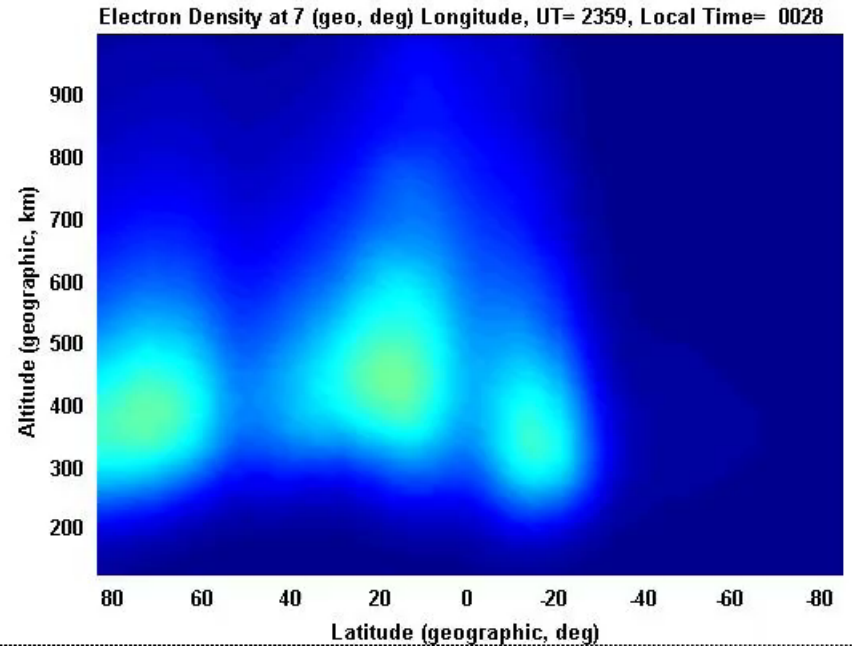
Feature Extraction is a Key Step in Data Mining and Machine Learning





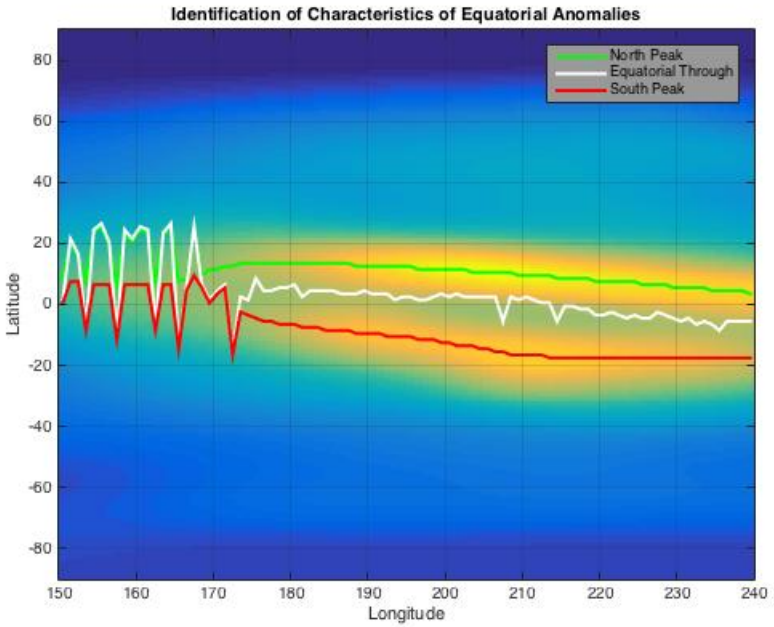
- Fingerprint recognition is one of the most mature machine learning application.
- Machine entire print leads to errors due to variability in prints.
- Features such as ridge ending, short ridges and bifurcation are more robust.

Physics Based Models Are Useful in Identifying Features in GIM



- Ion shower near equatorial region known as equatorial anomaly is related to electric field, thermosphere wind and solar radiation strength.
- The signature on a GIM is the most recognizable feature in a VTEC map.

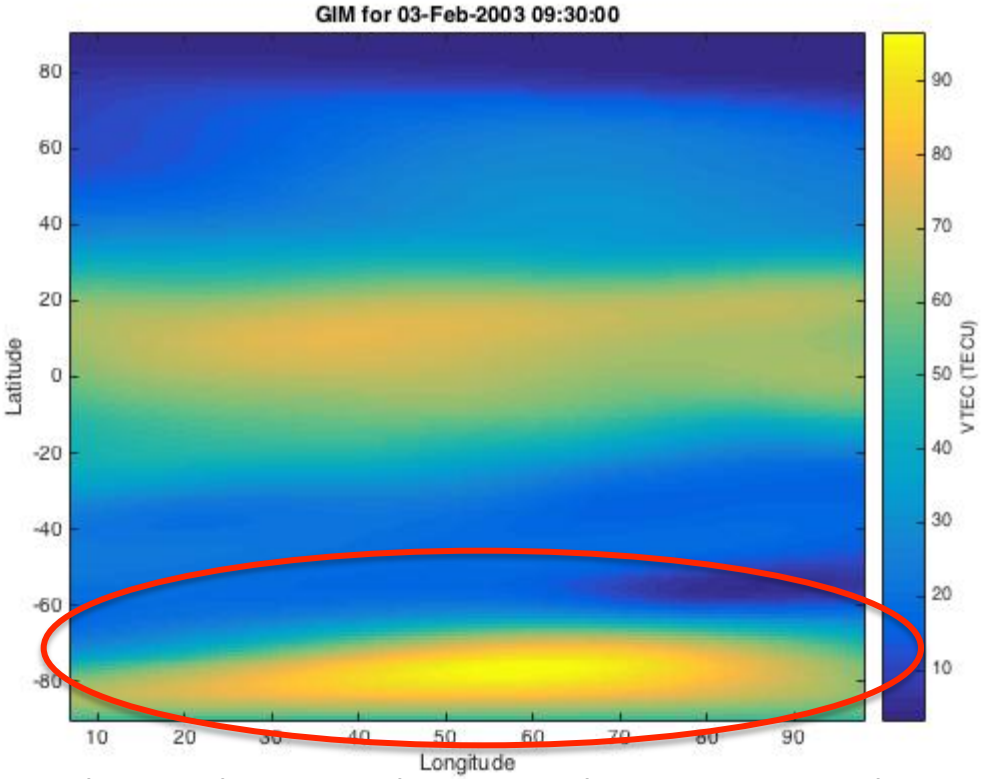
Equatorial Anomaly Can be Characterized Systematically



Automated algorithm is developed to identify the trough of VTEC near equatorial region between 11:00 and 18:00 local time.

- Key characteristics of the VTEC in equatorial region are measured.
  - Width of gap region between peaks in north and south of the equator.
  - VTEC difference between north and south equatorial peaks

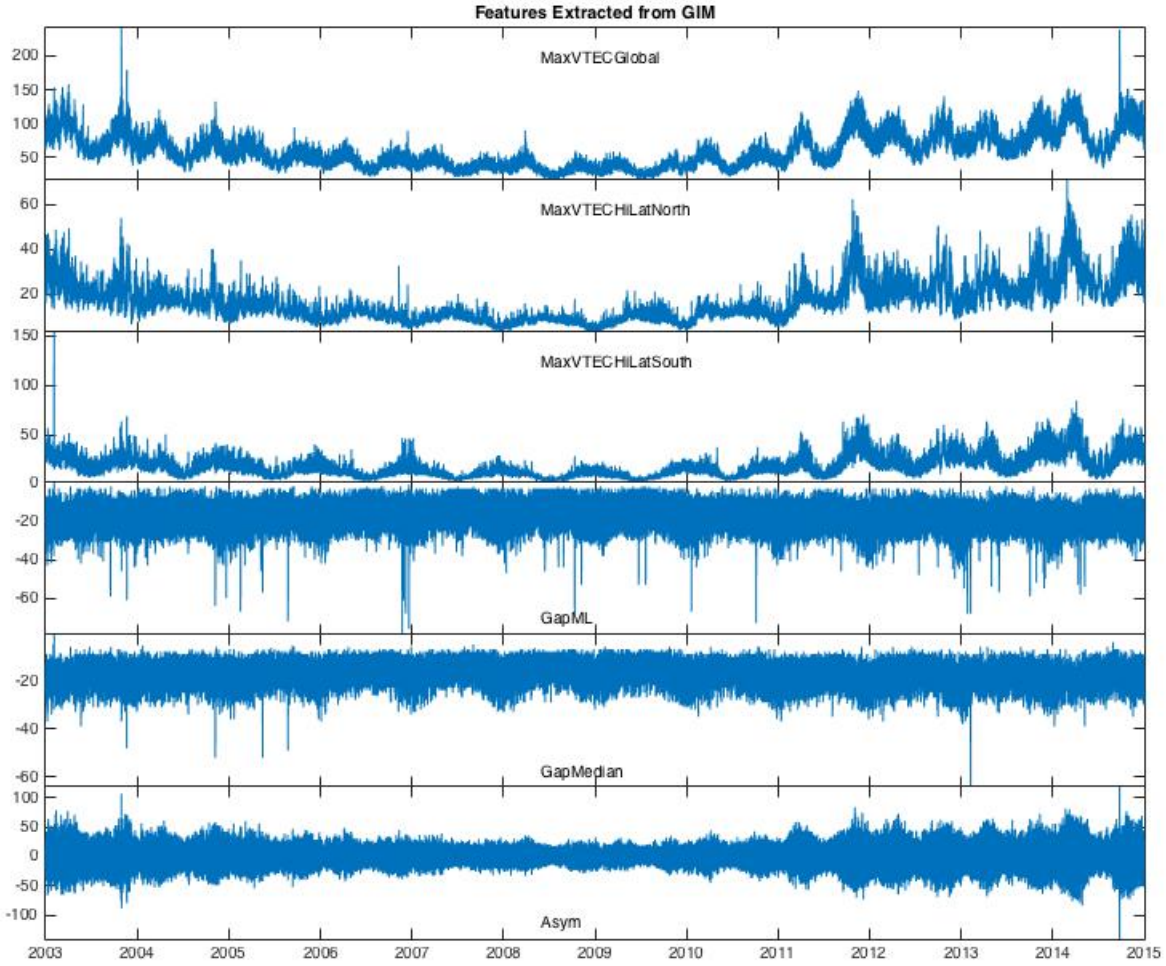
Global or Polar Region Peak VTEC Values Are Significant



Elevated VTEC values in polar region may be connected to strong geomagnetic and electrical fields disturbances.

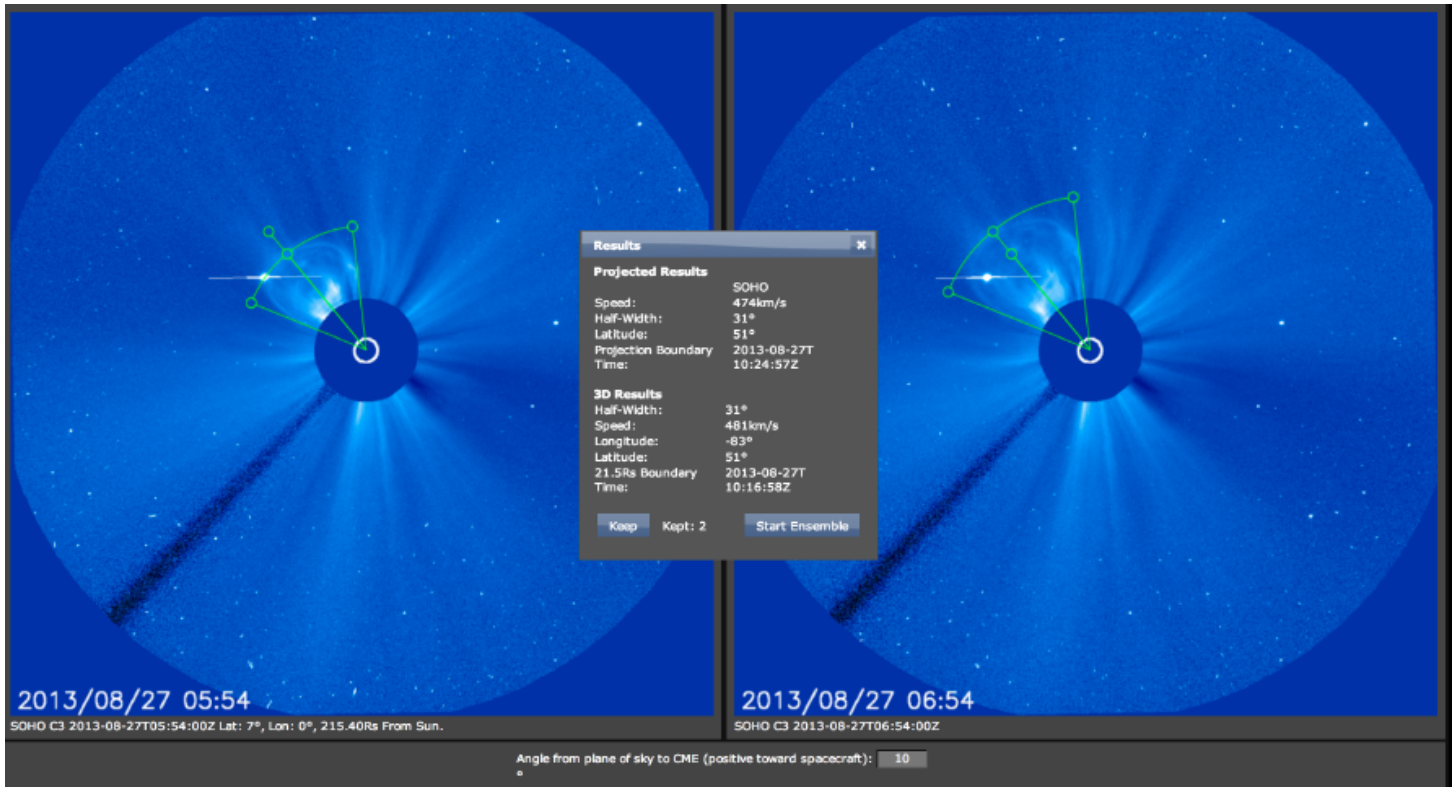
- Separated records of VTEC peaks for different regions reflect understanding of different physical phenomenon in these places.

Long-Term Historical Records of Key Features Are Essential

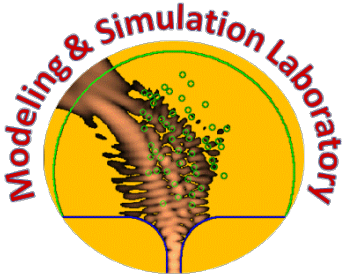


- Compilation and extraction of key features from historical data require expertise in space physics.
- Preparation of historical records helps make precise definition of features.

Feature Extraction Tools Are Already Developed for Coronagraph Analysis

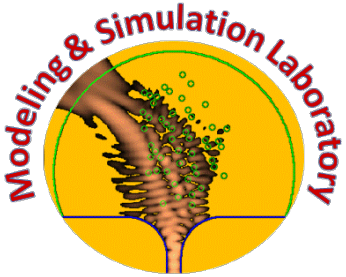


- StereoCAT and similar tools are routinely used to extract CME speed and other features.



Extreme Conditions Are Often Defined by Their Unusualness

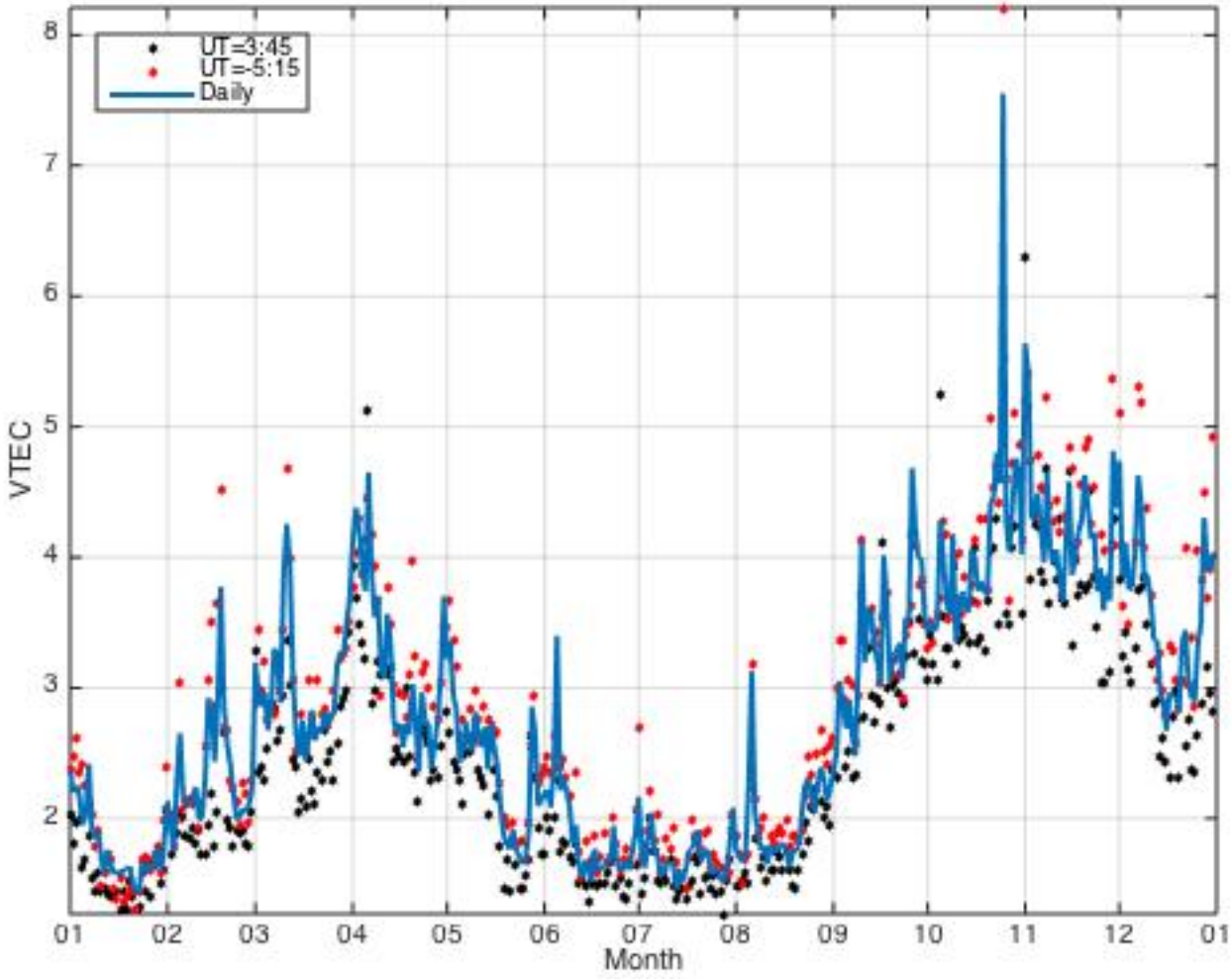
- In analyzing effects of strong solar or IMF disturbances on thermo-ionosphere it is common to identify the effects by comparing ionosphere measurements during the storm to those before and after.
  - Precise definition of thermo-ionosphere anomalies may not exist.
- Characterizing unusualness requires unbiased comparison to other measurement points.



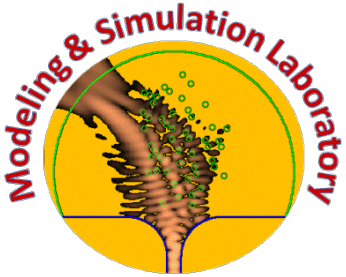
Identification and Detection of Outliers Are Parts of Machine Learning

- Unsupervised machine learning can be effective in discover useful features.
- Clusters in data records are important clues.
  - Different physically meaningful metrics can be used to measure the "distance" between data points.
  - Distance to nearest *n* neighboring points, Cluster Radius, provides an indication to the "unpopularity" of a data point.
  - The number of neighbors within a radius *r,* Popularity, represents similar information.

Computation of Cluster Radius Requires Intensive Processing

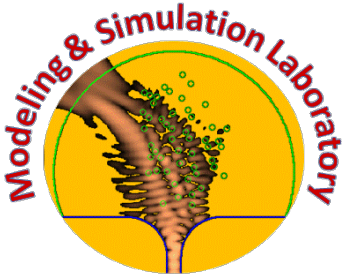


- Comparing daily GIM to that of m=365 prior days for 12 years required weeks or processing.
- Data compression techniques have the potential to significantly speed up processing.



Large Number of Metrics Are Used to Identify Relevant Features

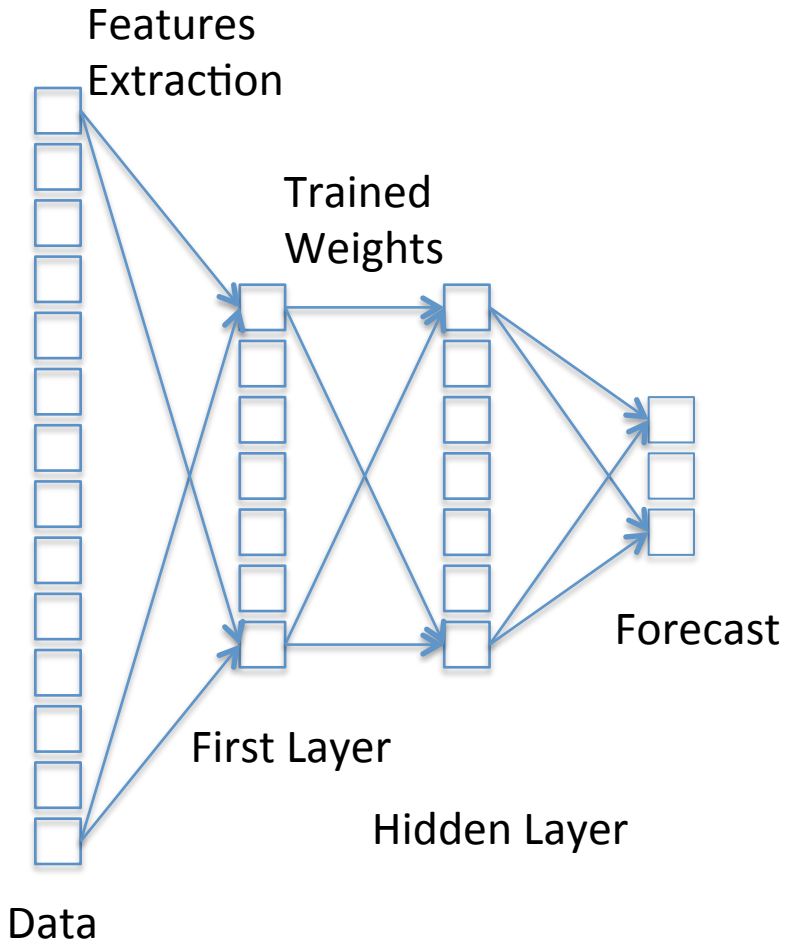
- All proposed metrics are motivated to capture physically meaningful aspects of the state of ionosphere.
  - Absolute VTEC difference;
  - Difference in latitude gradients;
  - Difference in key local time or latitude regions.
- A total of 12 different metrics are used to characterize the unusualness of a VTEC map.
  - Identification of the most useful metrics may lead to reduction in the number of metrics and optimization.



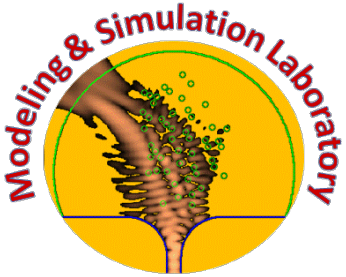
Outline

- Vision and Objective of the Test-bed
  - Paradigm and metrics of forecasting.
- Historical Database
  - Raw data and quality control of space data.
  - Ionosphere data and feature extraction.
- Regression Analyses
  - Forecasting strategies based on correlation and auto-correlations.
  - Training and validation of regression based forecast.
- Support and continued development of the test-bed
  - Potential data sources and other forecasting strategies.

Machine Learning Consists of Mining Historical Data for Useful Clues

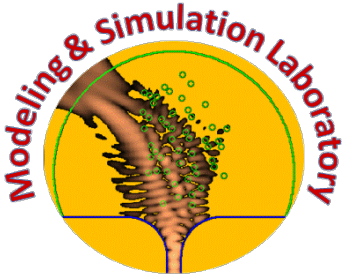


- Many different techniques exist.
  - Different representations of relationship between input and output variables.
  - Different optimization criteria.
  - Different technique for searching for optimal parameters.
- All are regression analyses?



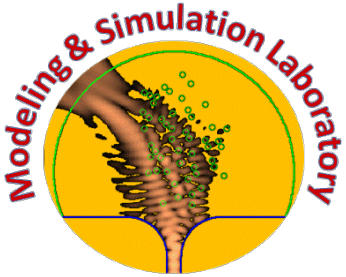
Correlations Between Historical Data Are Fundamental

- Data can be used for thermo-ionosphere forecast for time t_1 produced at time t_0 include all measurements of solar, IMF, geomagnetic and thermo-ionosphere up to time t_{-1}.
  - The delay $d=t_0-t_{-1}$ is referring to ask data latency which may be different for different measurement.
  - The difference $f=t_1-t_0$ is the advance for the forecast. Zero advance is referred to as nowcast.
- Correlations in historical data of two measurements separated by time difference of $d+f$ are useful.



Fading Memory in Space Weather Helps Limit Correlations in Time

- Current set of space and ionosphere data include 128 variables at 3 hour resolution.
- Using a 5 day interval, there are 40x128=6450 components for correlation analyses.
- Precise definition of training data, forecast data and validation data must be established.
  - Training data: data available before t_{-1} used to find optimal regression parameters to be used for a forecast.
  - Forecast data: data available at time t_{-1} and within fading memory interval.

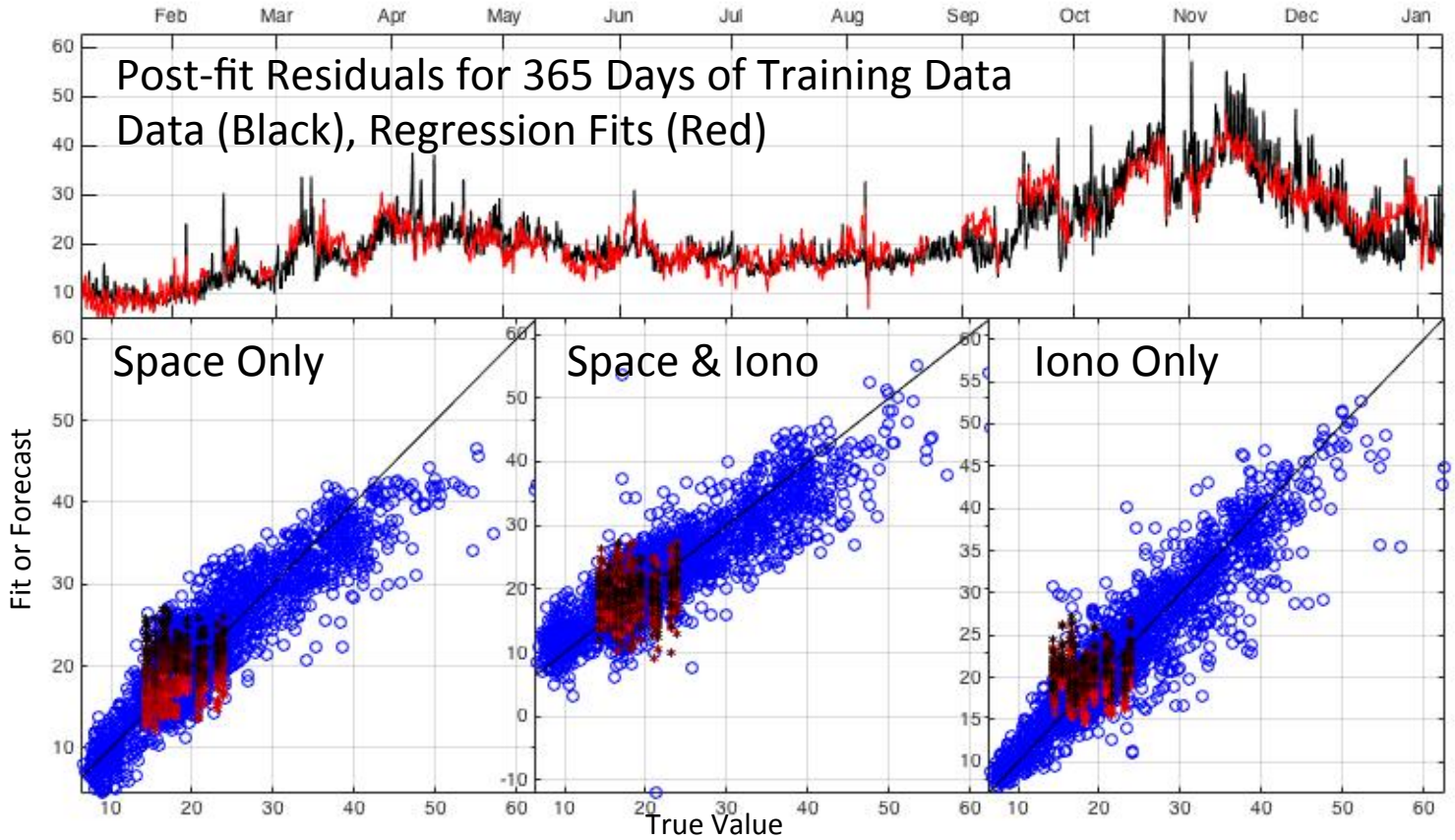


Validation of Forecast Strategy Must Respect Data Availability Constraints

- All regression analyses for finding optimal forecast parameters must only use data available before t_{-1}.
- Validation of forecast strategy consists of comparing the forecasted value $v_{forecat}$ to measured value v_{truth} for time t_1.
- Regression parameters can be updated as time t_1 is moving forward.
  - Instead of a fixed “empirical” forecast formula, continuous regression analyses are able to catch new trend in data.

Large Number of Forecasting Strategies Can be Explored

Regression residual for 1 day forecast compared to validation of multiday forecasts



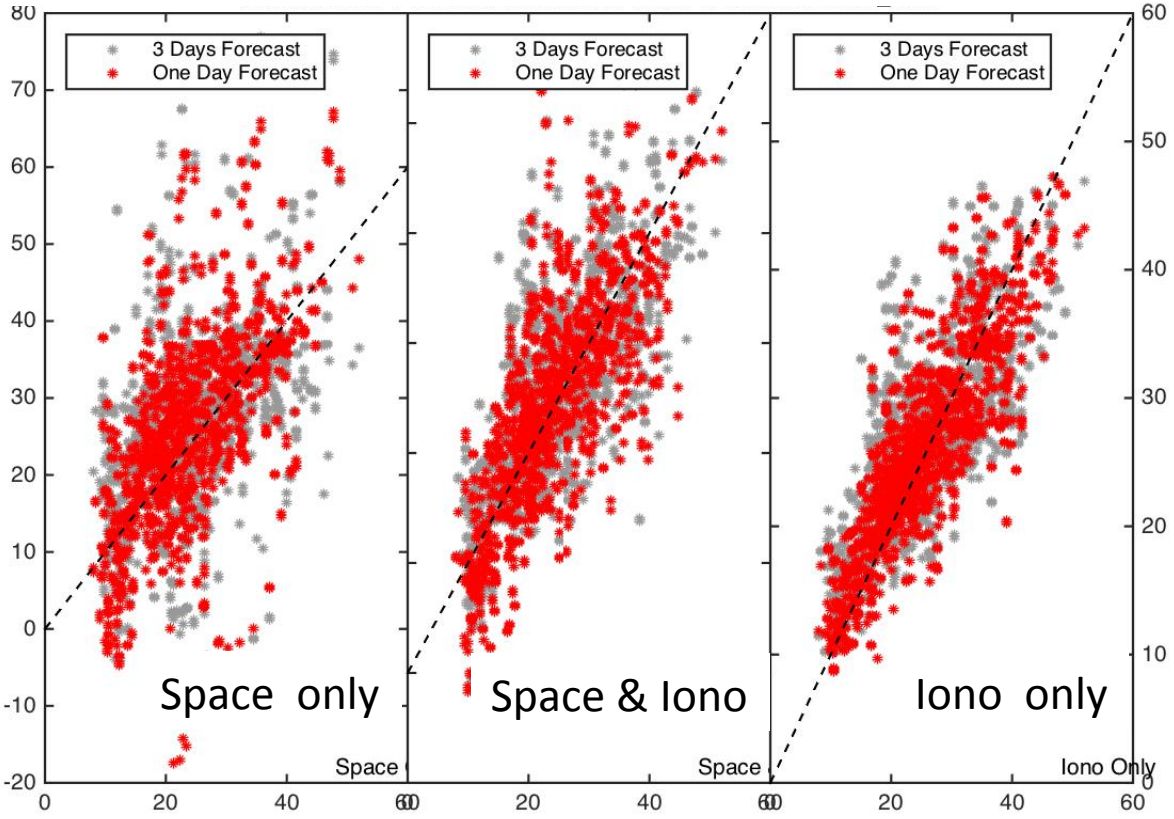
Training Data (Blue Circle) and Validation Data (Different Shade of Red)

Systematic Validation of Linear Regression Based Forecast is Performed

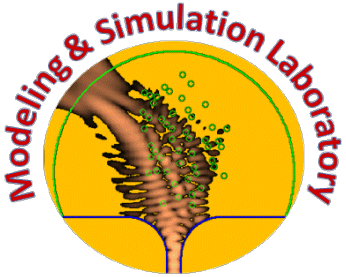
- Validation data for forecast of all 169 variable with *d +f* ranging from 0 to 4 days are computed for 2011.
- Large number of automated analyses are performed and the results are being analyzed.
  - Initial results show encouraging performances.
  - The results are preliminary and require further and close examination.
  - Statistical analyses of the validation data in term of True-Positive Rate vs. False-Positive rate must be performed.

Increase of Number of Parameter May Not Lead to Improvement

Forecast Performance for Maximum VTEC in Southern Hemisphere High Latitude Region for All 2011

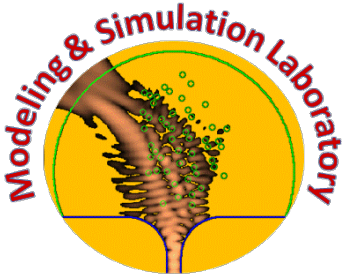


- Using only ionosphere data seems produce best results.
- Increased degree of freedom also increase instability.



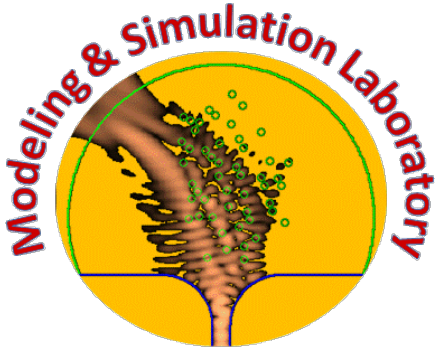
Outline

- Vision and Objective of the Test-bed
  - Paradigm and metrics of forecasting.
- Historical Database
  - Raw data and quality control of space data.
  - Ionosphere data and feature extraction.
- Regression Analyses
  - Forecasting strategies based on correlation and auto-correlations.
  - Training and validation of regression based forecast.
- Support and continued development of the test-bed
  - Potential data sources and other forecasting strategies.



SWFT Can Only Succeed With Community Participation

- Talent from all participants in space weather research are needed.
  - Data provider must supply quality controlled data.
  - Space physicists can help identify relevant features.
  - People affected space weather must contribute in the definition of value of a forecast.
  - Young data scientists can use their knowledge to contribute in the discoveries in space weather.
- We should seek broad community guidance and acceptance of the SWFT.

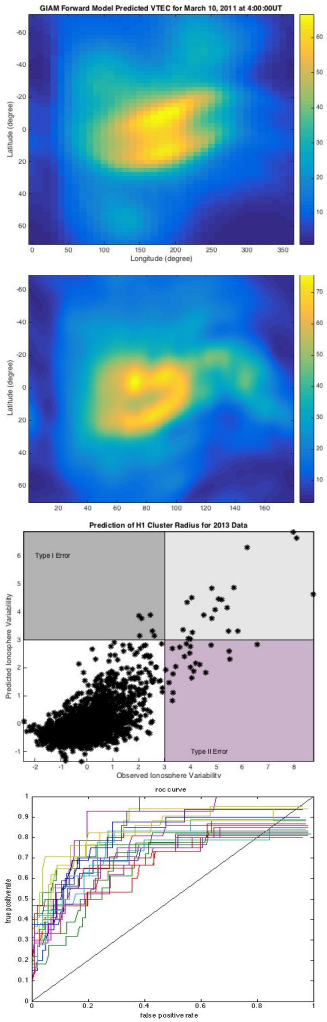


Statistical Techniques for Forecasting Ionosphere Anomalies Using Solar and Space Observations

Chunming Wang

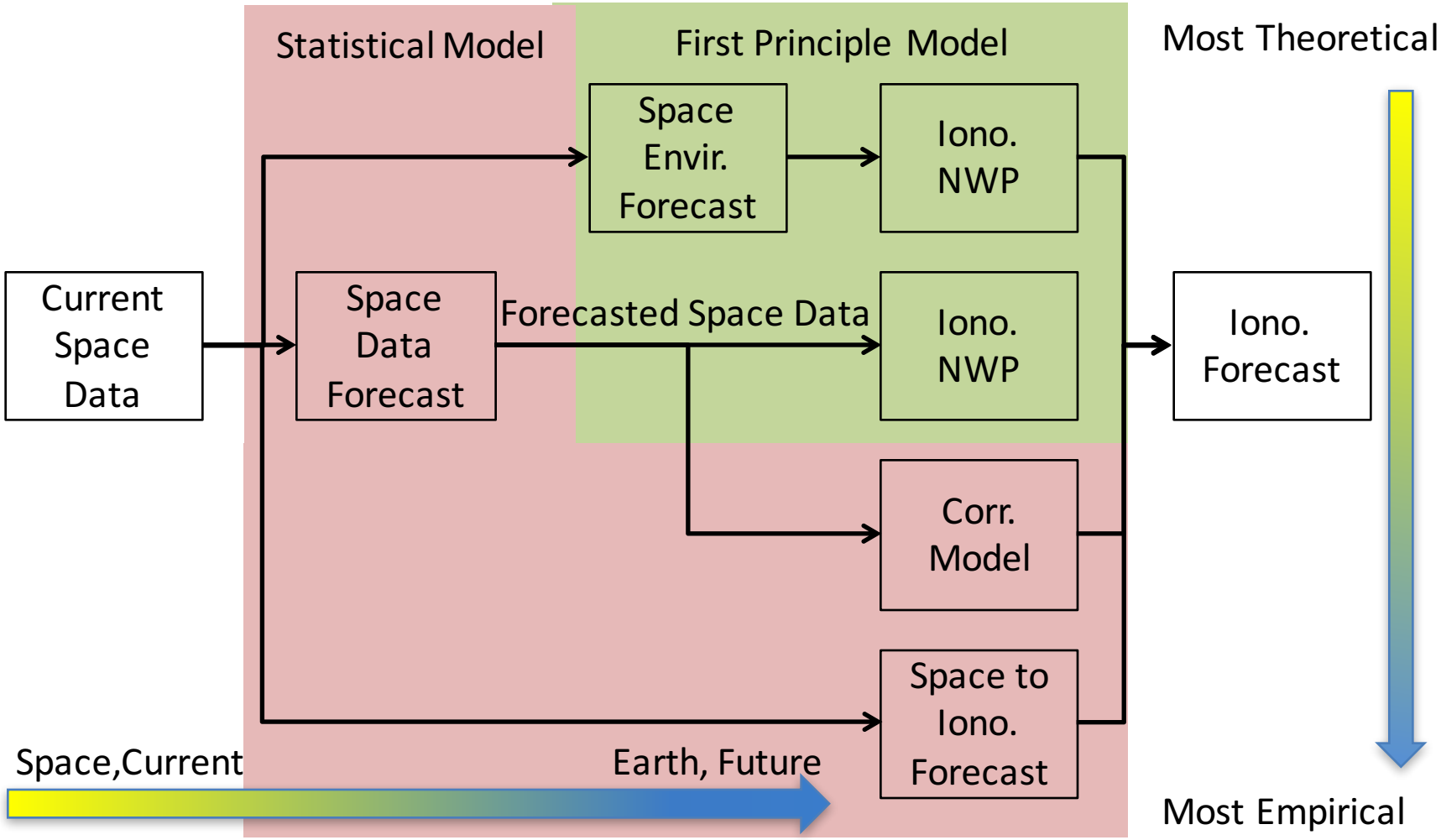
Modeling & Simulation Laboratory
Department of Mathematics
University of Southern California

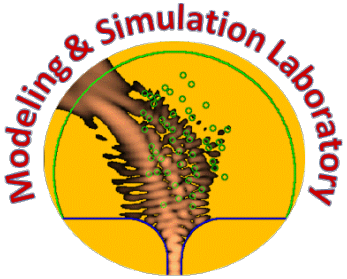
Statistical Techniques Have Always Been Crucial Tools for Space Weather Forecast



- Calibration of first-principle based model often requires data currently unavailable.
- Statistics based system identification, regression and machine learning techniques help discover empirical connections between solar and space observation and ionosphere anomalies.
- Forecast comes with statistical confidences derived from historical data.

Statistical Methods Provide Key Bridges to Fill Gaps Between Models





Our Efforts Cover Several Important and Practical Aspects of Ionosphere Forecast

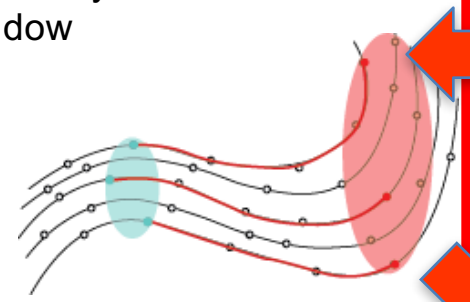
- Discovering hidden dynamics in space weather observations for forecasting solar and space environment (Kayo Ide, Eugenia Kalnay, Erin Lynch and Surja Sharma,University of Maryland, College Park)
- Statistical characterization of ionosphere anomalies and their connections to space environment anomalies via regression analysis (G. Rosen & C.W,USC)
- Machine learning approach for direct forecast of ionosphere anomalies using solar and space weather observations(G. Rosen & C.W,USC)

Analyses of Historical Solar and Ionosphere Data Revealed Important Correlations

Forecast Space Observation Using Data Driven Dynamic Model

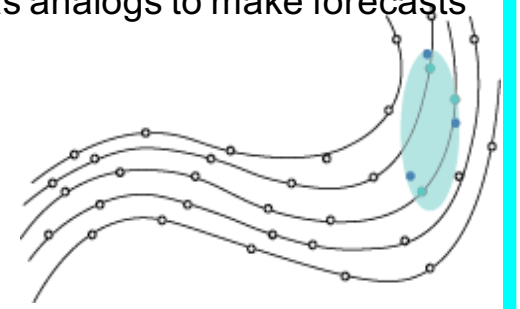
"Model" Forecast

Use a dense data set of points on the attractor (model) to advance NN analysis ensemble to the end of the analysis window



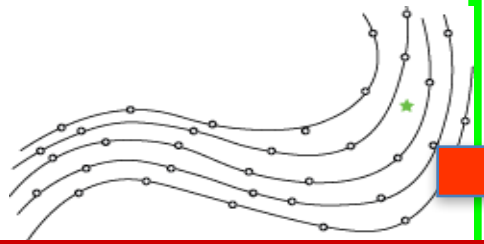
Nearest Neighbors (NN)

Locate nearest neighbors of analysis ensemble members to serve as analogs to make forecasts



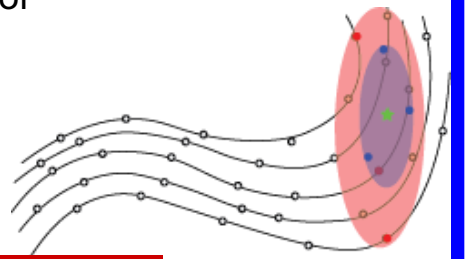
Observations

Observations of a single variable (i.e. the AL index) become multivariate when embedded



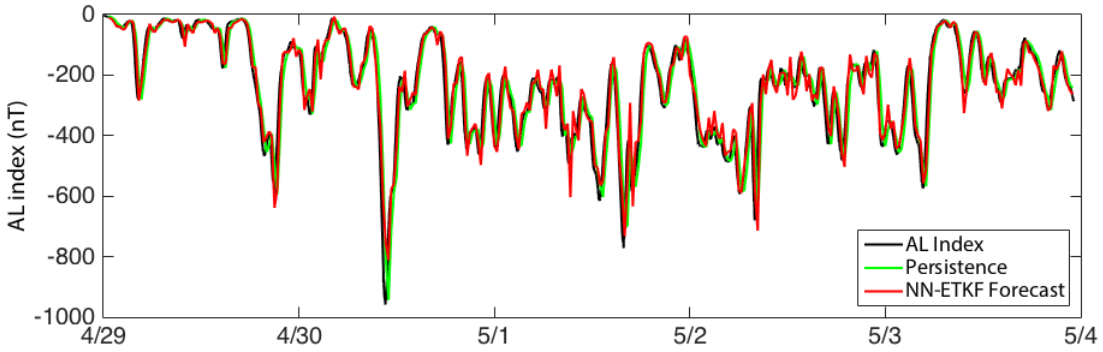
Analysis

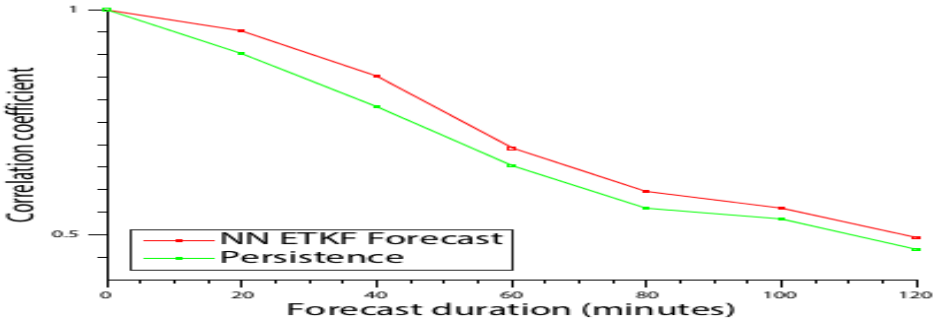
Analysis ensemble members computed using the ETKF are the best estimates of the true state, but do not lie on the attractor

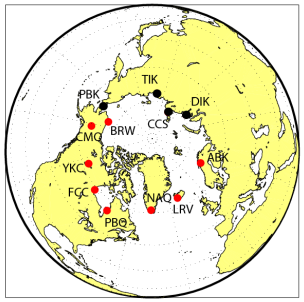


Kayo Ide, Eugenia Kalnay, Erin Lynch and Surja Sharma, University of Maryland, College Park

Empirical Technique Enables Forecasts of Magnetic Field Variations at Ground Stations



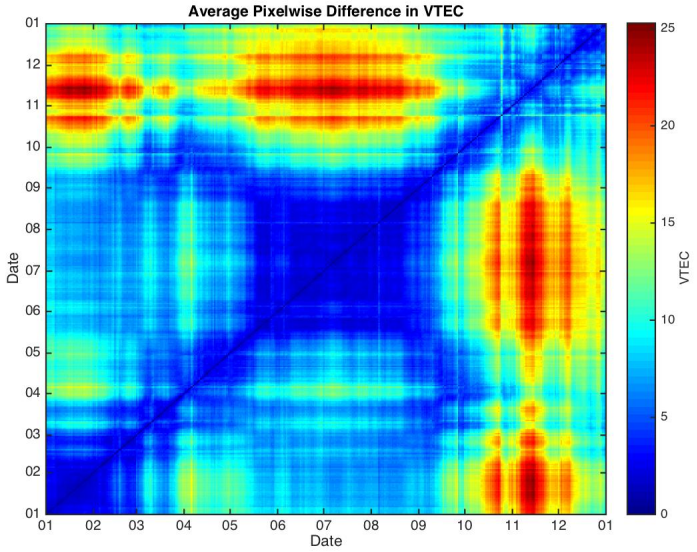




- Twelve ground-based magnetometer stations provide data to the World Data Center for the construction of the AL indices
- Eight stations have publicly available data.
- 40 minute forecasts of four of the fluctuations in magnetic field readings of four of the stations during the April 2011 HSS event
- Forecasts of the eight stations with available data were performed simultaneously

Kayo Ide, Eugenia Kalnay, Erin Lynch and Surja Sharma, University of Maryland, College Park

Unambiguous Characterization of Anomalies in the Ionosphere is Essential



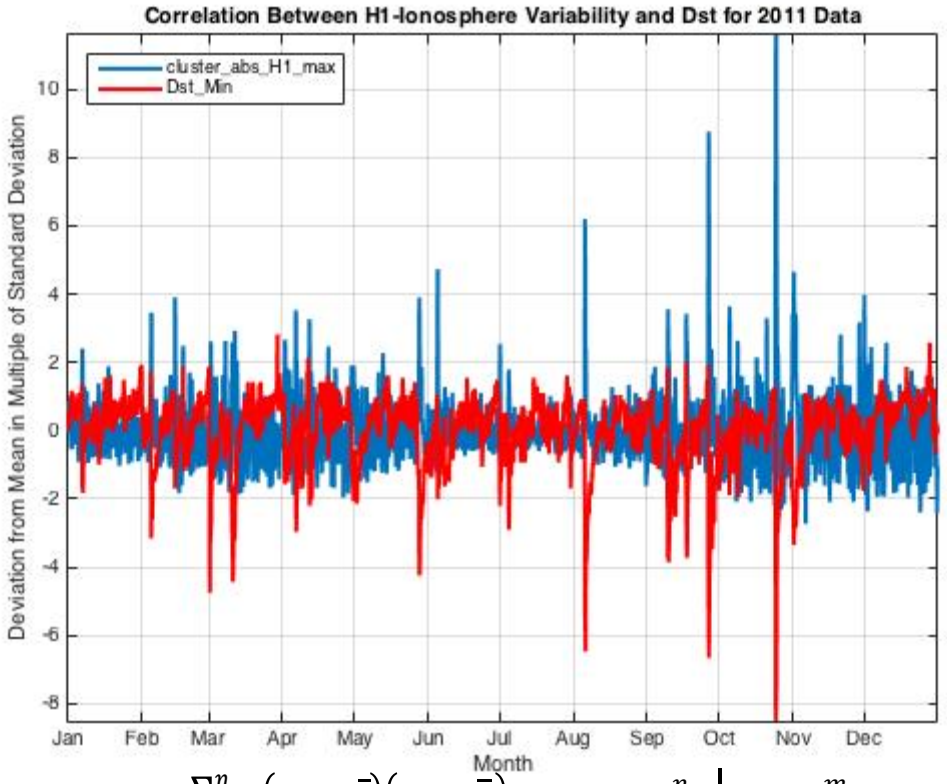
- 16 different metrics are used to measure the difference between two days of GIM/VTEC maps.
- Radius of the cluster of nearest *n* neighbors is used to characterize the unusualness of GIM/VTEC maps for a given day.
- Significant correlation is observed to identified solar and space anomalies.

Extensive Solar, IMF and Other Space Weather Indices Are Analyzed

- Using solar, IMF and other space weather data directly avoids delays of intermediate analyses.
- Identification of apparent correlation between space weather data and ionosphere variability indices is inherently useful.
- Analyses used 2011-2014 data.

- Solar and space Data
  - Bartels Number,
  - Bartels Phase
  - Kp,Ap,Cp,
  - SunSpot,F107
  - Dst
  - Bx, By,Bz
  - Vx, Vy, Vz
  - Proton density
  - Temperature
  - Flow pressure
  - Ae, Al, Au
  - Pcn

Correlation Coefficients Measure the Coincidence of Changes in Two Variables



$$co(x,y) = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{(n-1)\sigma_x\sigma_y}$$

Correlation Coefficient

$$\min_x \sum_{k=1}^{n} \left| y_k - \sum_{j=1}^{m} \alpha_j x_{j,k} \right|^2$$

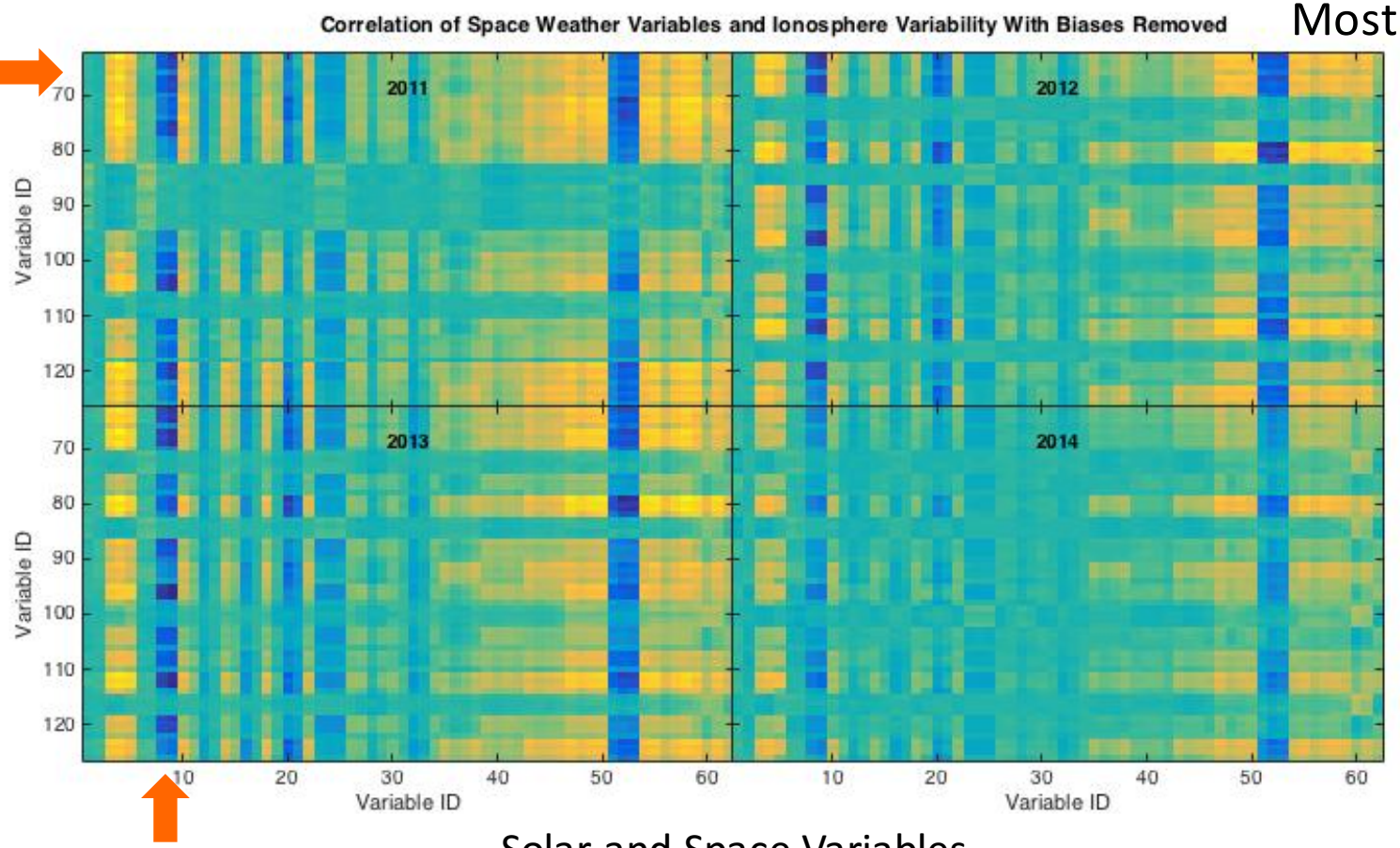
Regression Problem

- Substantial correlation is observed between the minimum value of Dst and many ionosphere variability metrics.
- Negative correlation is expected since negative values of Dst correspond to disturbance conditions.

Statistically Significant Correlations Are Observed Between Space and Ionosphere

#67 cluster_abs_H1_max

Ionosphere Variability



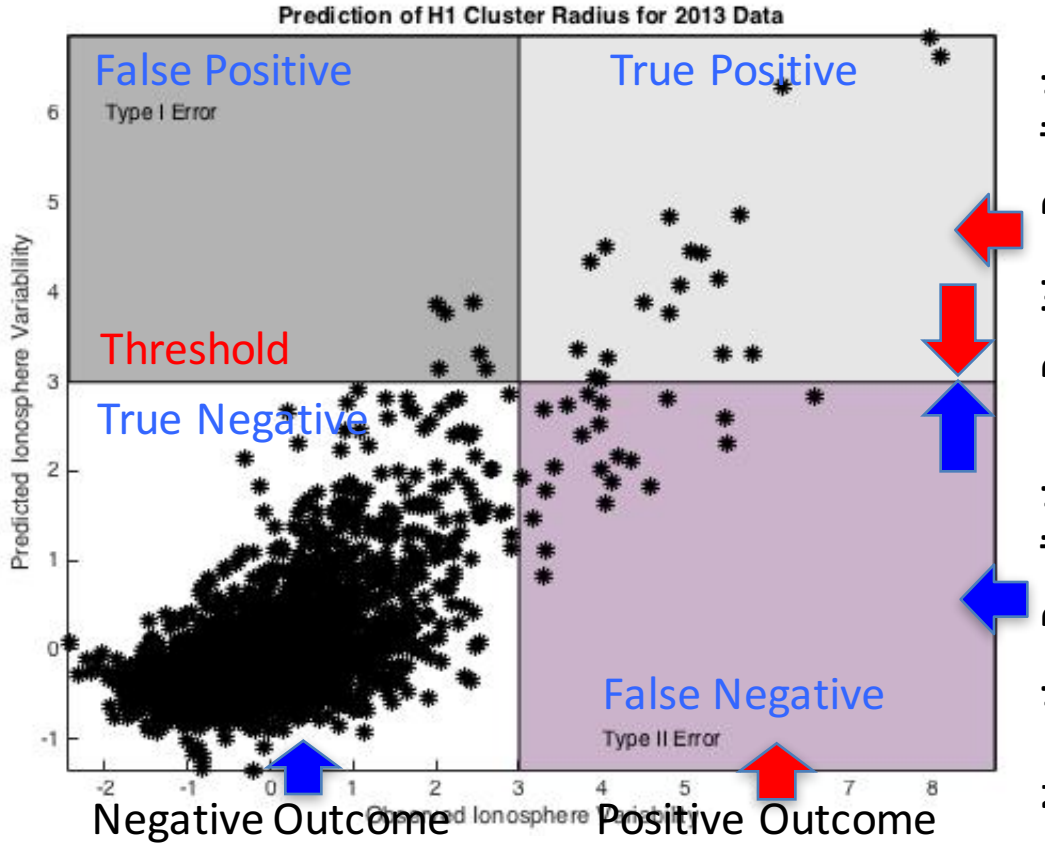
#9 Dst_Min

Solar and Space Variables

Most Significant

Ap
Dst
Kp
Pcn
AU
AE
AL
CP
FPres
Bx
By
Bz

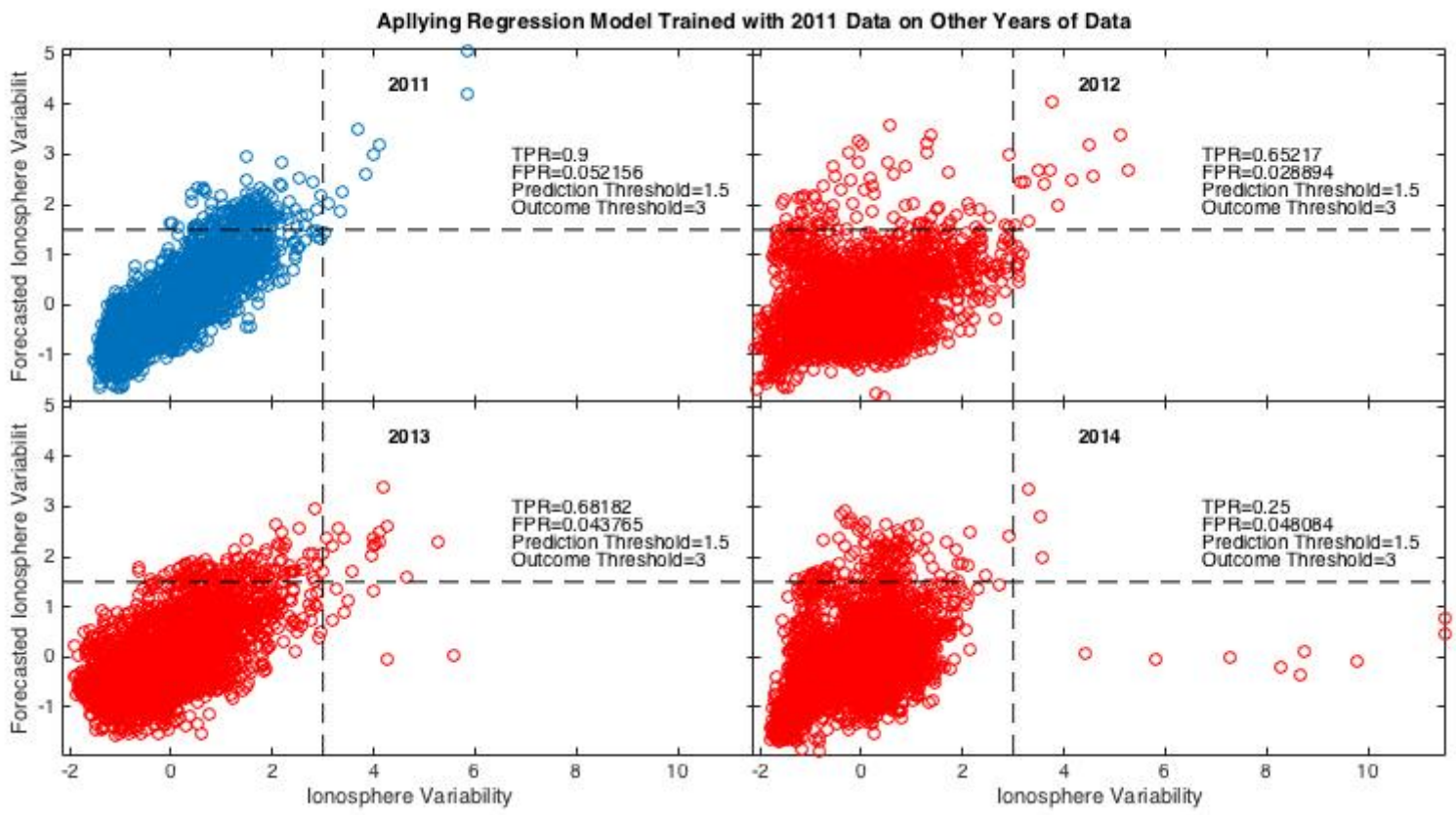
Regression Analyses Allows Quantification of Forecasting Errors



$$FPR = \frac{\text{Netagtive Outcome and Positive Prediction}}{\text{Negative Outcome}} \qquad TPR = \frac{\text{Positive Outcome and Prediction}}{\text{Positive Outcome}}$$

- Regression analyses produce similar quality of prediction of ionosphere variability.
- Many components of space variables are strongly correlated.
- Redundancy of space variables leads to instability of regression coefficients.

Regression Model Trained with 2011 Shows Reasonable Predictability

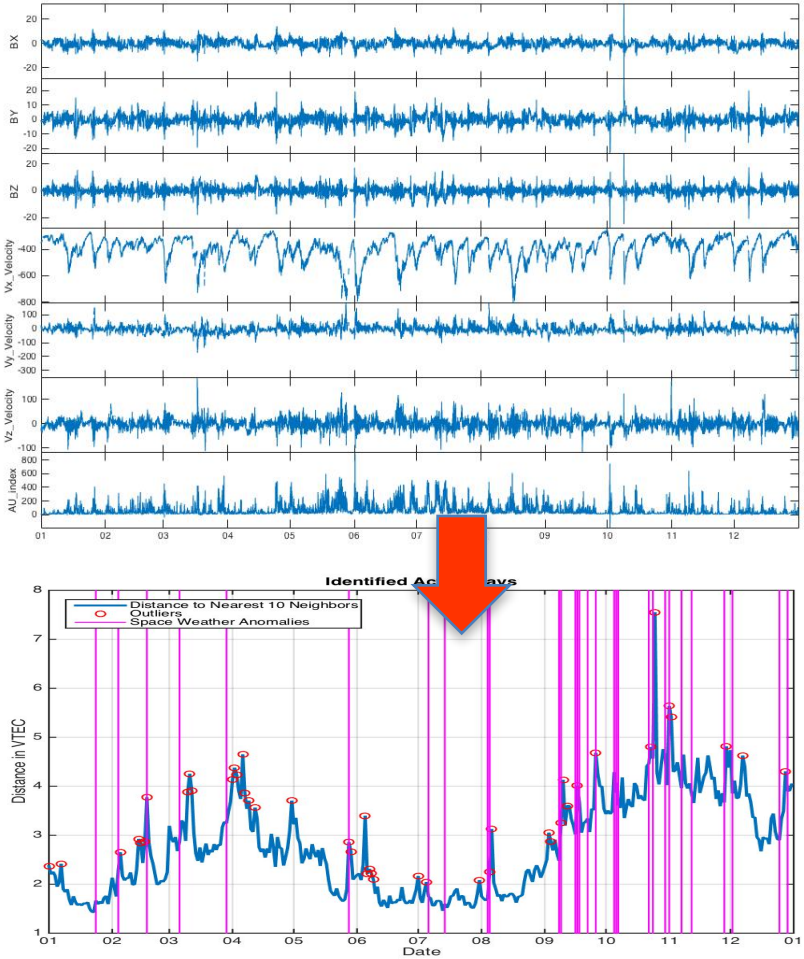


- There are outliers in 2014 data that need to be examined.

Most Significant

Ap
Dst
Kp
Pcn
AU
AE
AL
CP
Flow Pres
Bx
By
Bz

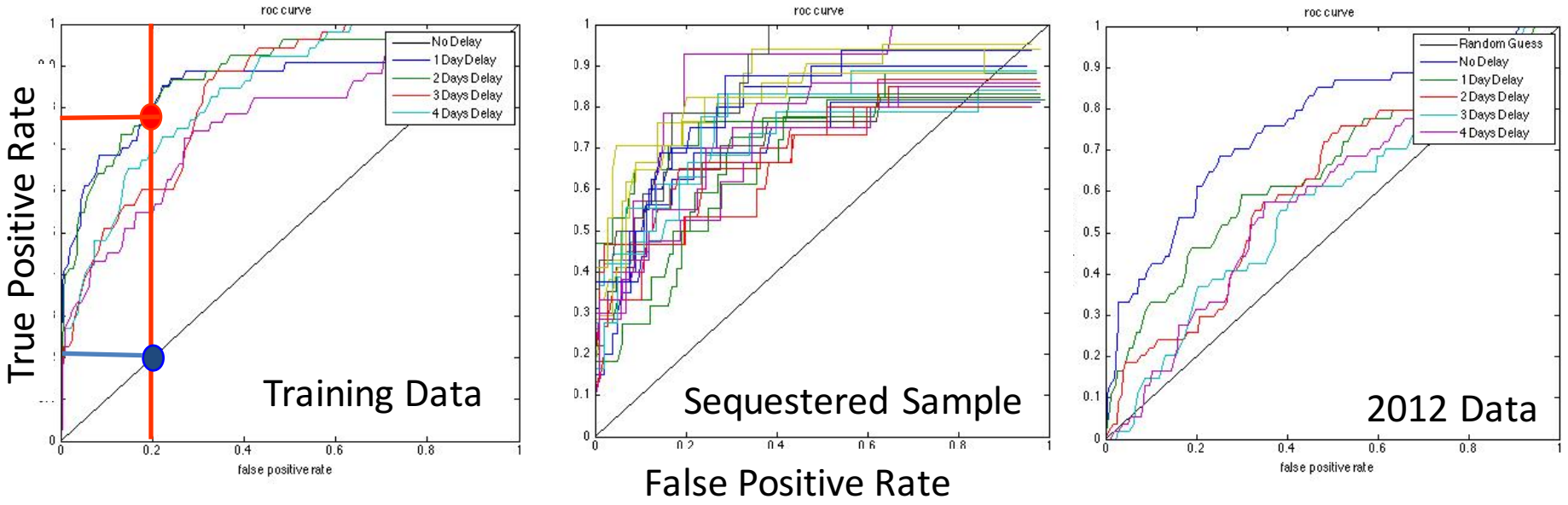
Using Machine Learning Techniques to Construct Anomaly Forecasting Model



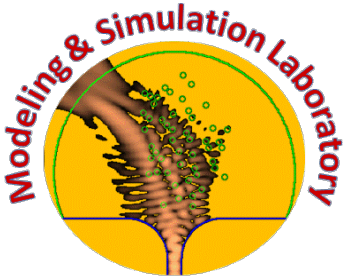
Use *logistic regression* to build a *binary classifier*

- **Dependent or Response Variable**: GIM TEC maps *labeled* 1 if an outlier and *labeled* 0 if not an outlier.
- **Independent or Predictor Variables:** a selection from 60 observed space weather parameters or features.
- Use ***Binomial Logistic Regression*** to fit data from 2011.

Validation Against Historical Data Allows Quantification of Statistical Confidence



- Based on ***Receiver Operating Characteristic (ROC) curves***; True Positive Rate (TPR) vs False Positive Rate (FPR) for various cut-off thresholds
- ***Cross validation*** on 2011 data and on 2012 data



Future Research Efforts

- Expanding the Kalman filter approach for the forecast of solar-wind and Al indices
  - Solar-wind forecast can be used by first principle models
- Develop and evaluate systems for forecasting of ionosphere disturbance using solar and space observation with historical data
  - Derive cluster radius using only data available prior to prediction
  - Update regression analysis continuously
  - Develop forecast skill measurements