Statistical Learning Methods for Big Data Analysis and Predictive Algorithm Development

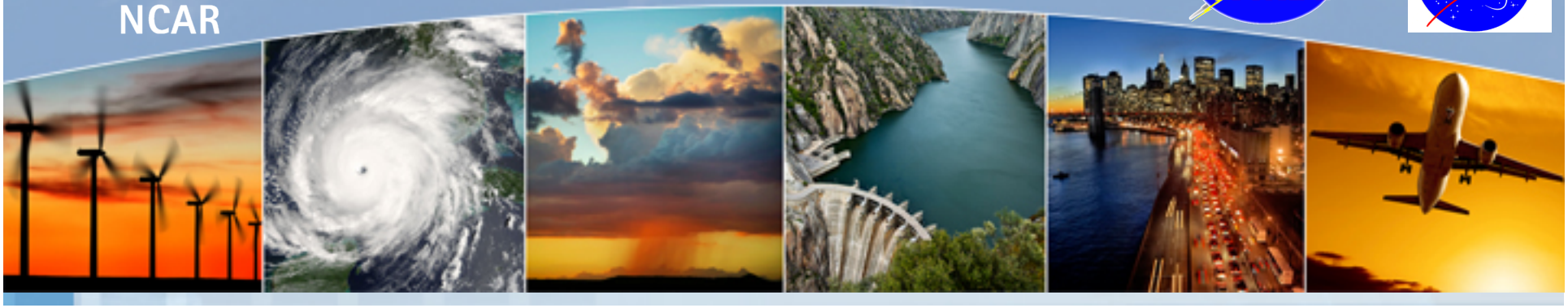
John K. Williams, David Ahijevych, Gary Blackburn,
Jason Craig and Greg Meymaris
NCAR Research Applications Laboratory

SEA Software Engineering Conference

Boulder, CO

April 1, 2013

NCAR



Outline

- Big data
- Statistical learning
- Sample applications from CIDU
- Empirical modeling for decision support
- Use case: Aviation turbulence diagnosis
- Use case: Convective storm nowcasting
- Big data and statistical learning challenges
- Resources and opportunities

Big Data

- "Big data" ≈ data too large to handle easily on a single server or using traditional techniques
- E.g., atmospheric sciences data: rapidly ballooning observations (e.g., radar, satellites, sensor networks), NWP models, climate models, ensemble data, etc.
- Improved management and exploitation recognized as key to advances in government-sponsored research and private industry
- Challenges include:
  - Limiting number of formats
  - Consistent, adequate metadata
  - Ontologies for data discovery
  - Accessing, using and visualizing data
  - Server-side processing and distributed storage

Big Data

- In early 2012, the federal government announced Big Data Research and Development Initiative, which unified and expanded efforts in numerous departments
- Big data examples:
  - FAA "4-D data cube" for real-time weather and other information
  - NASA Earth Exchange (https://c3.nasa.gov/nex/)
  - NSF EarthCube, evolving via a community-oriented iterative process and grants
  - Solicitations for developing Big Data initiatives
- *Adequately exploiting Big Data requires developing and applying appropriate statistical learning techniques for knowledge discovery and user-relevant predictions.*

Statistical Learning

NCAR

- A collection of automated or semi-automated techniques for discovering previously unknown patterns in data, including relationships that can be used for prediction of user-relevant quantities
- A.k.a. data mining, machine learning, knowledge discovery, etc.

Statistical Learning Family Tree

Mathematics

"Soft" computing

Statistics

Optimization Theory

Computational Intelligence

Machine Learning

Data Mining

Expert Systems

Supervised Learning

Unsupervised learning

Fuzzy Logic

Decision Trees

PR Poly. Reg-ression

SVM Support Vect. Machine

DT Decision Trees

GA Genetic Algorithm

SA Simulated Annealing

RL Reinforc. Learning

NBC Naïve Bayes Class

ANN Art. Neural Networks

KNN *k-Nearest Neighbors*

RF Random Forest

PCA Principal Comp. Anal.

SOM Self-organ. map

FCM Fuzzy *c*-means clust.

Statistical Learning Family Tree

Mathematics

"Soft" computing

Statistics

Optimization Theory

Computational Intelligence

Machine Learning

Data Mining

Expert Systems

Supervised Learning

Unsupervised learning

Fuzzy Logic

Decision Trees

PR
Poly. Reg-ression

SVM
Support Vect. Machine

DT
Decision Trees

GA
Genetic Algorithm

SA
Simulated Annealing

RL
Reinforc. Learning

NBC
Naïve Bayes Class

ANN
Art. Neural Networks

KNN
k-Nearest Neighbors

RF
Random Forest

PCA
Principal Comp. Anal.

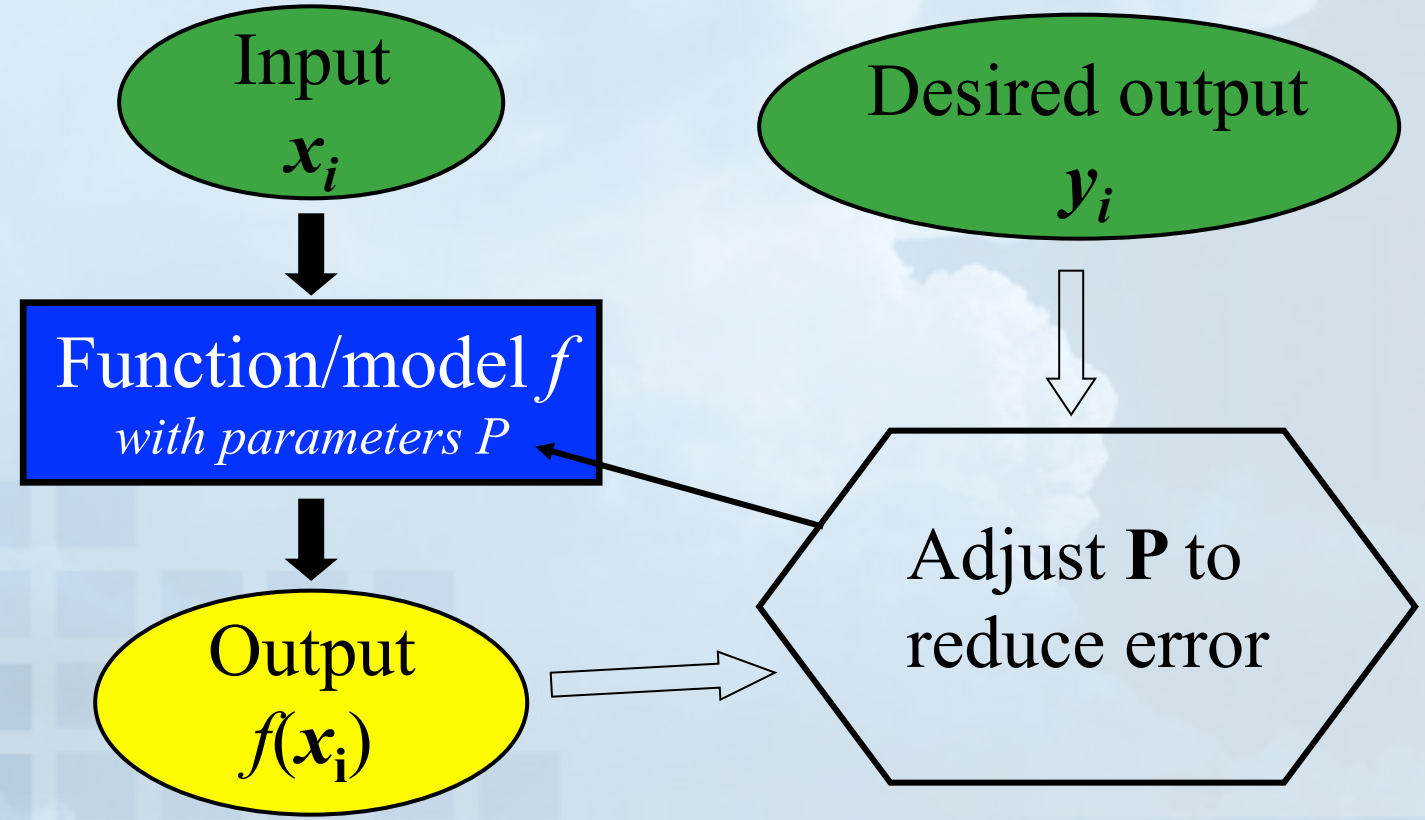
SOM
Self-organ. map

FCM
Fuzzy *c*-means clust.

Supervised Learning for Empirical Modeling

NCAR

- Uses "training" data: many labeled examples ($\mathbf{x}_i$, $\mathbf{y}_i$)
  - regression ($\mathbf{y}_i$ in R^n) or classification ($\mathbf{y}_i$ in a discrete set)
- Learn model parameters to reduce some error function
- If all goes well, model will "generalize" well to new data!



Linear Regression

- Linear model:

$$f(x) = m\,x + b$$

parameters

- *More generally, with a vector of predictors:*

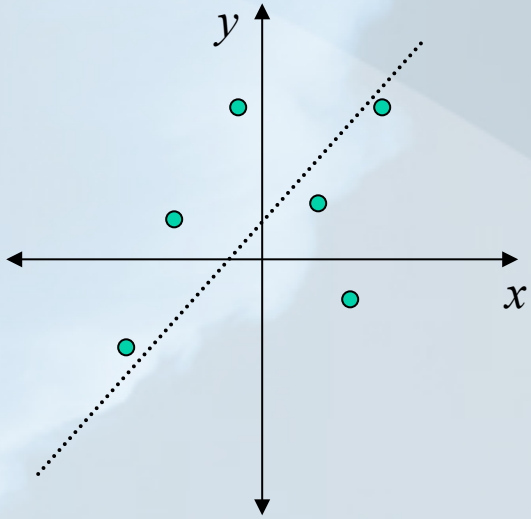
$$f(\mathbf{x}) = \mathrm{A}\,\mathbf{x} + \mathbf{b}$$

- Error function (sum of squared errors):

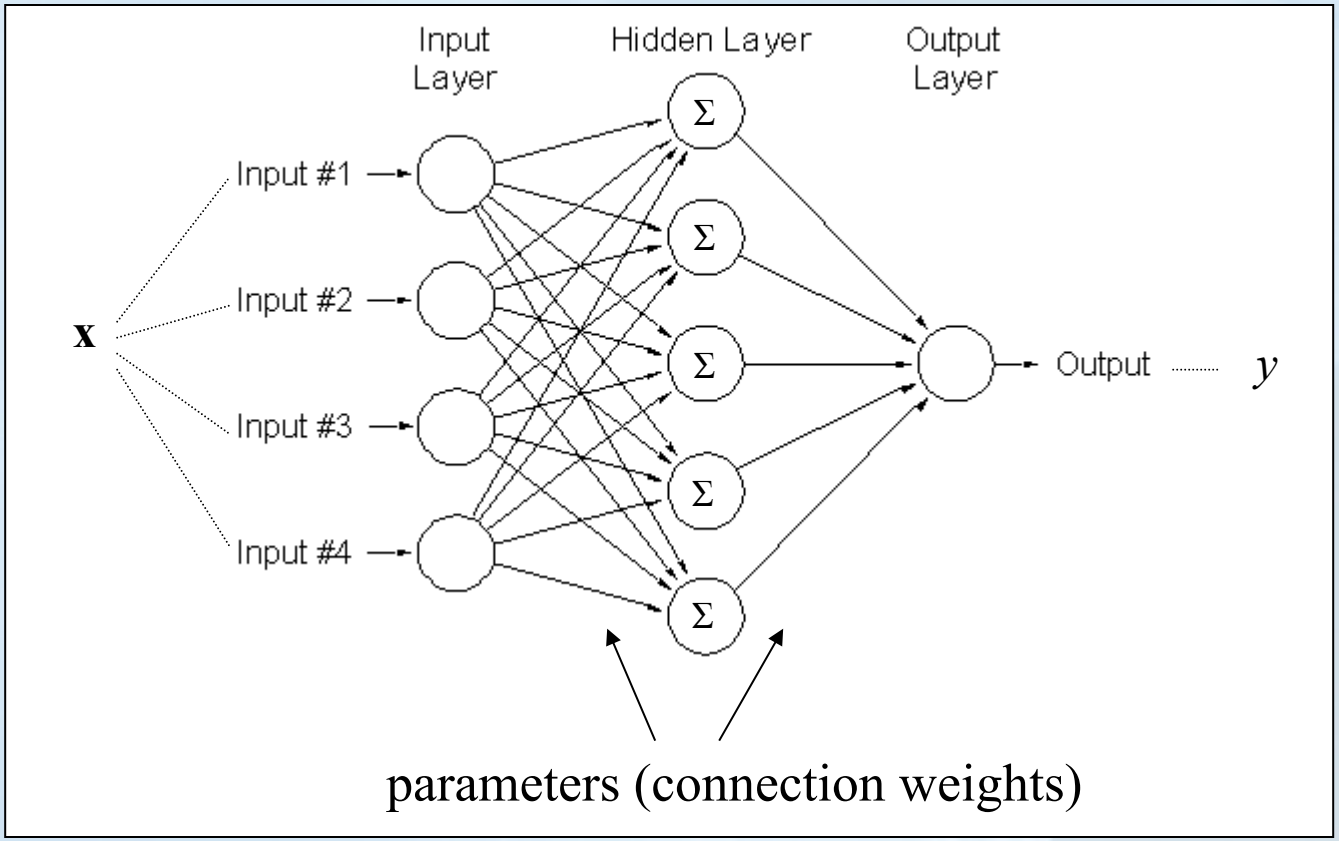
$$\Sigma_i \left(f(\mathbf{x}_i) - y_i\right)^2$$

learn m and b (or $\mathbf{A}$ and $\mathbf{b}$) to minimize this SSE

i	x_i	y_i
1	2	2
2	5	6
3	-3	2
4	-6	-4
5	-1	6
6	3	-1



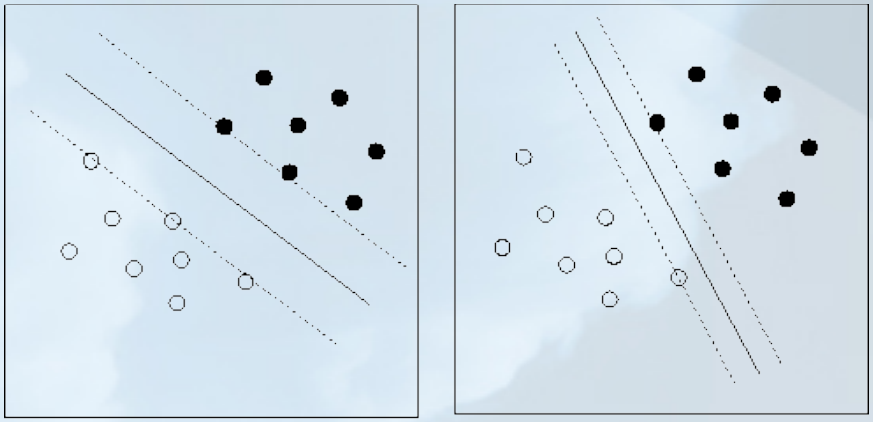
Artificial Neural Network (ANN)



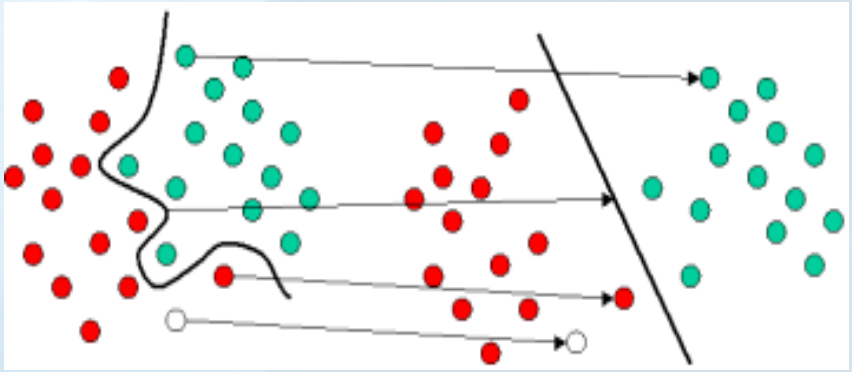
- Use "back-propagation" to adjust weights based on output errors (gradient descent)
- Any function can be approximated with large enough hidden layer (but you can also fit the noise in your data!)
- But trained models are not human-readable

Support Vector Machine

- An efficient way to learn classification using a maximum (soft) margin
- Maps data space into an infinite-dimensional "feature" space where it performs linear classification

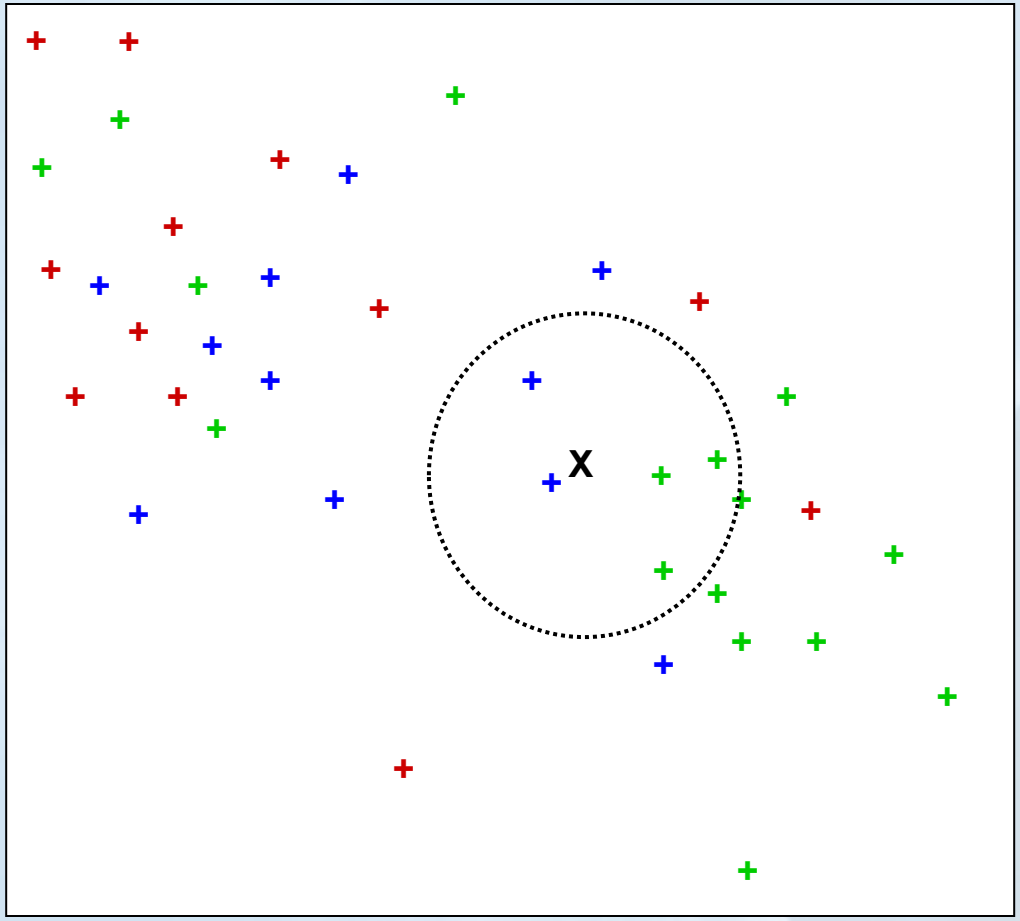


Linear classification: The larger margin separation at left has superior generalization



SVM: maps to a higher-dimensional space where linear classification is possible

k-Nearest Neighbor Classifier

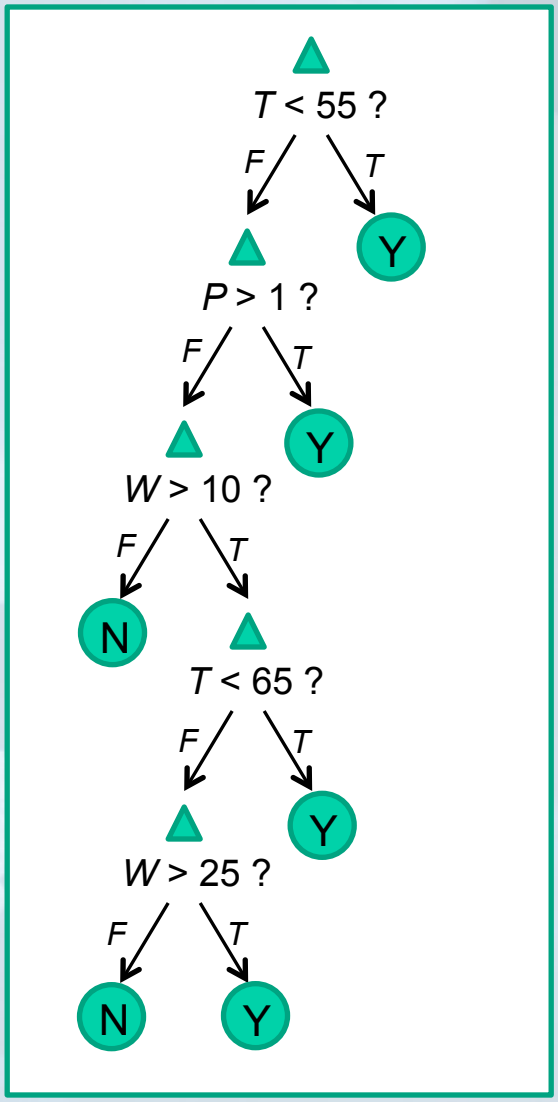


- For positive integer *k* classify a new point (at **X**) based on "consensus" of the *k* nearest labeled examples
- Not really a "model," more an inference method

Decision Tree

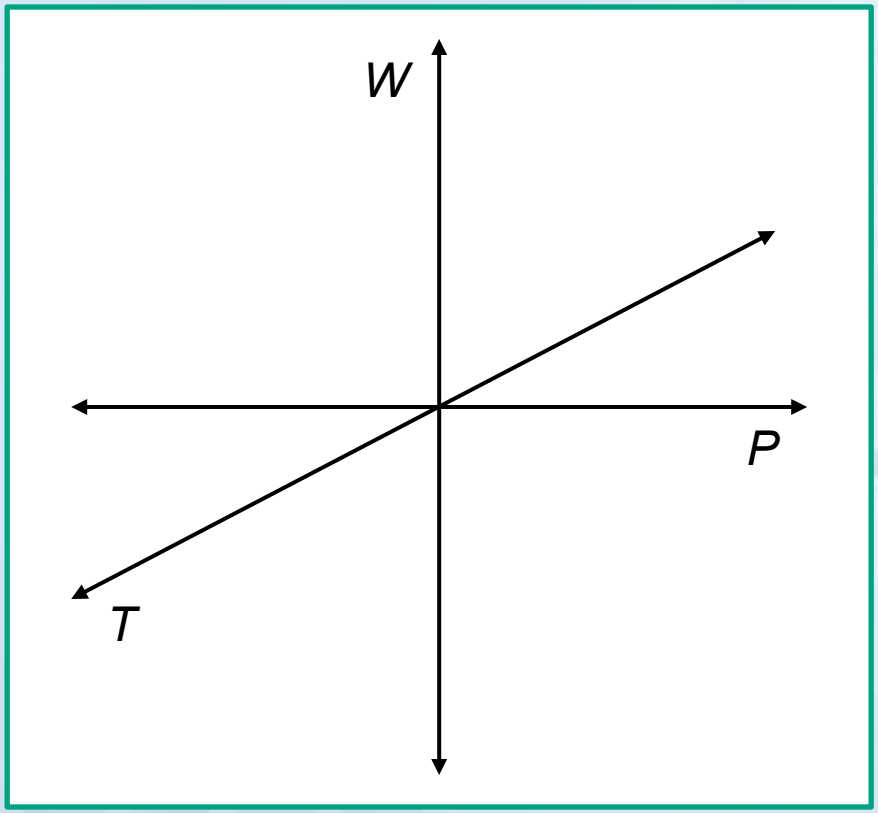
- E.g., “Should I take a coat?”

Temp *T* (°F)	Precip *P* (mm/hr)	Wind speed *W* (mi/hr)	Wanted Coat?
50	0	5	Y
63	2.5	2	Y
90	0	18	N
67	0	30	Y
72	0.5	12	N
42	5	8	Y
83	0	22	N
	. . .		

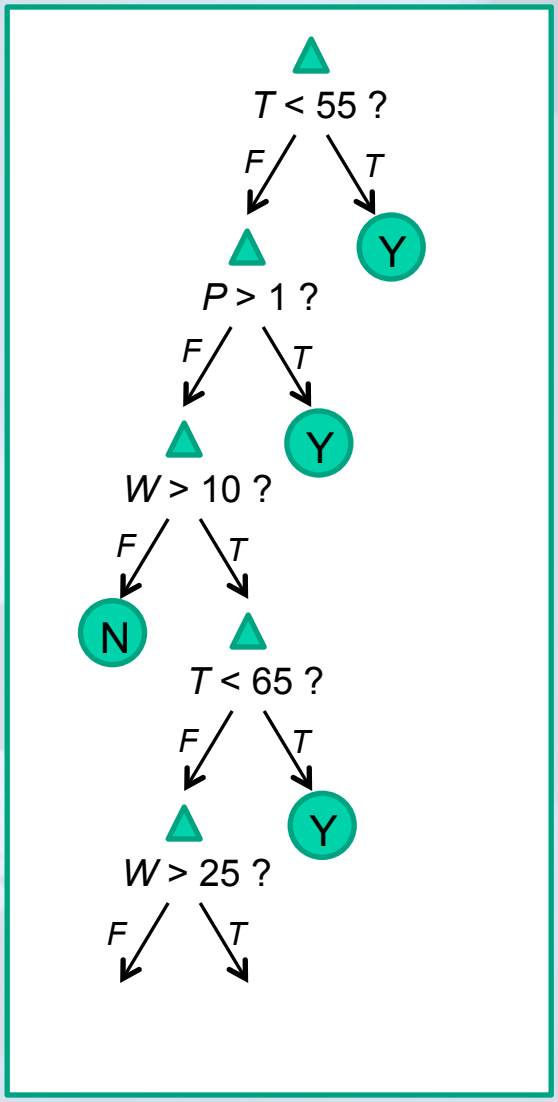


Decision Tree (cont.)

- E.g., “Should I take a coat?”



The decision tree assigns a category (Y or N) to every point in *T, P, W* space



Decision Tree (cont.)

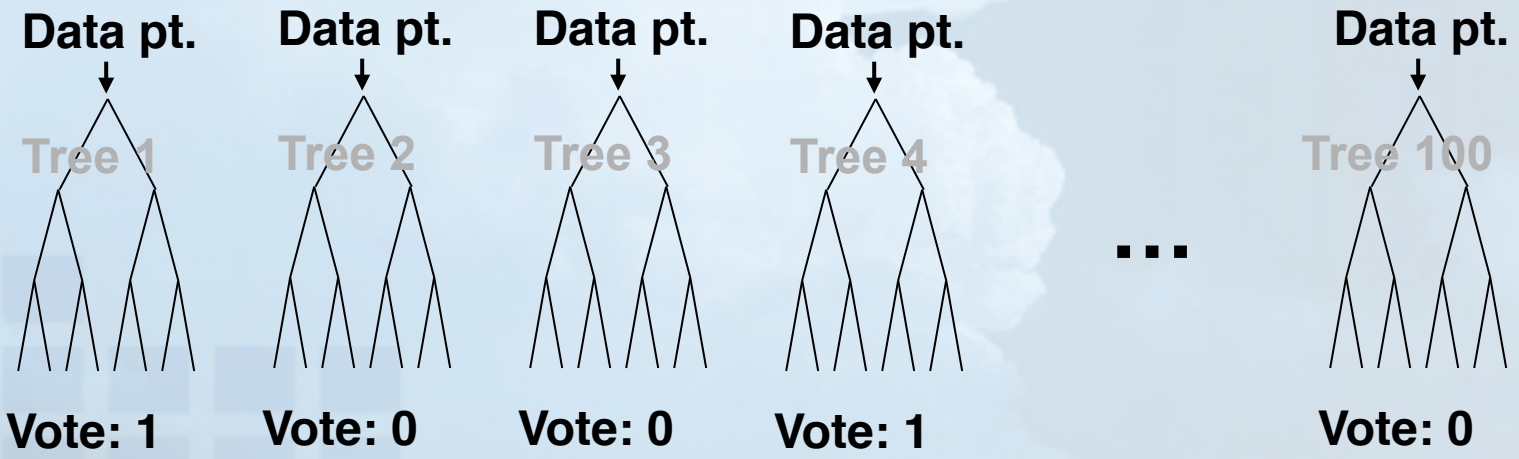
- Can represent any function that is constant over “boxes” in predictor space, a good approximation even to complex, nonlinear functions
- Classification (discrete values) or regression (continuous values)
- Decision trees can be “grown” automatically from a “training” set of labeled data by recursively choosing the “most informative” split at each node
- Trees are human-readable and are relatively straightforward to interpret
- But in order to generalize well to an independent testing dataset, trees must be limited in size or “pruned”
  - Balancing simplicity and performance can be an “art”

Random Forest (RF)

- Produces a *collection* of decision trees using predictor variables and associated class labels (for classification) or values (for regression)
  - Each tree is based on a random subset of data and predictor variables, making it (somewhat) independent from others
  - The ensemble is less prone to over-fitting and other problems of individual DTs, and generally performs better
- During training, random forests produce estimates of predictor "importance" that can help in selecting predictor variables and understanding how they contribute
  - Roughly speaking, importance of a variable estimates the change in classification accuracy when its values are randomly permuted among instances
  - Variables that have been selected more frequently and have greater significance for the problem show greater importance

RF, continued

- A trained random forest provides an empirical model that can be applied to new data
  - the trees function as an "ensemble of experts", voting on the predicted classification for each new data point
  - the vote distribution may be calibrated to create a non-parametric (makes no assumptions about functional form), probabilistic empirical predictive model



E.g., 40 votes for "0", 60 votes for "1"

Sample Statistical Learning Applications from CIDU-2012 (papers)

- Species Distribution Modeling and Prediction
- Estimation and Bias Correction of Aerosol Abundance using Data-driven Machine Learning and Remote Sensing
- Machine Learning Enhancement of Storm Scale Ensemble Precipitation Forecasts
- Hierarchical Structure of the Madden-Julian Oscillation in Infrared Brightness Temperature Revealed through Nonlinear Laplacian Spectral Analysis
- Time Series Change Detection using Segmentation: A Case Study for Land Cover Monitoring
- Importance of Vegetation Type in Forest Cover Estimation
- A Technique to Improve the Quality of Accumulated Fields
- EddyScan: A Physically Consistent Ocean Eddy Monitoring Application96
- A New Data Mining Framework for Forest Fire Mapping
- Learning Ensembles of Continuous Bayesian Networks: An Application to Rainfall Prediction
- Data Understanding using Semi-Supervised Clustering
- Mining Time-lagged Relationships in Spatio-Temporal Climate Data

Sample Statistical Learning Applications from CIDU-2012 (posters)

- A data-adaptive seasonal weather prediction system based on singular spectrum analysis and k-nearest neighbor algorithms
- Classification of Hand Written Weather Reports
- Optimizing the Use of Geostationary Satellite Data for Nowcasting Convective Initiation
- Feature-Based Verification for Wind Ramp Forecasts of the High-Resolution NCAR-Xcel WRF-RTFDDA
- Statistical analysis of intra-farm wind variation using wind turbine nacelle and meteorological tower wind observations
- Genetic Algorithms and Data Assimilation.
- The Dynamic Integrated ForeCast System
- Using a Random Forest to Predict Convective Initiation
- Time-series Outlier Detection by Clustering
- Self-Organizing Maps as a method to identify patterns and pattern change in regional climate models, with application to renewable energy resource assessment
- A Data Mining Approach to Data Fusion for Turbulence Diagnosis

RAL Focus: Technology Transfer

NCAR

Goal: Real-time decision support

Comprehensive products that

- synthesize relevant information
- reduce forecaster and user workload
- support user decisions

Example users

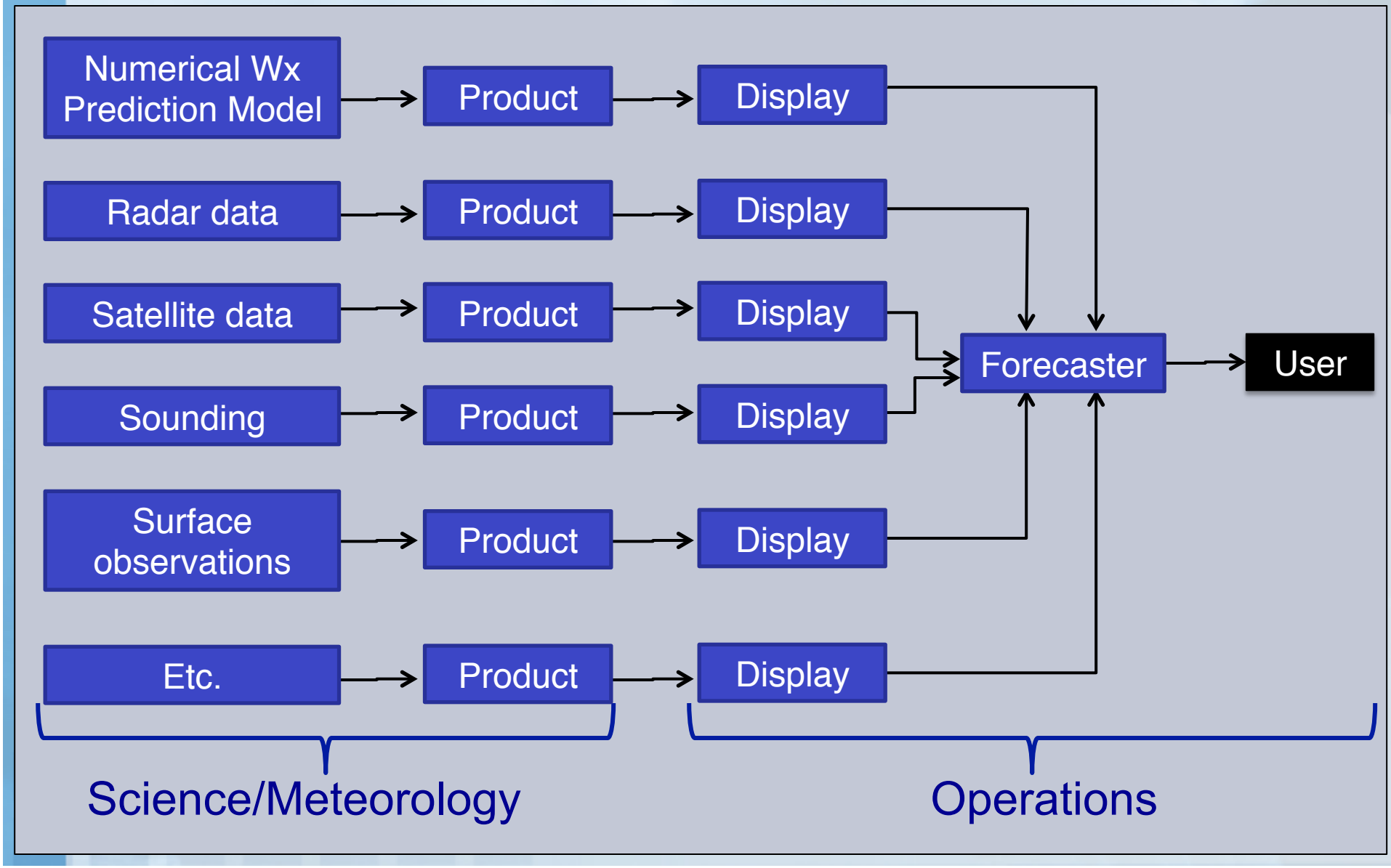
Aviation: Dispatchers, air traffic managers, pilots

Energy: Utilities, traders

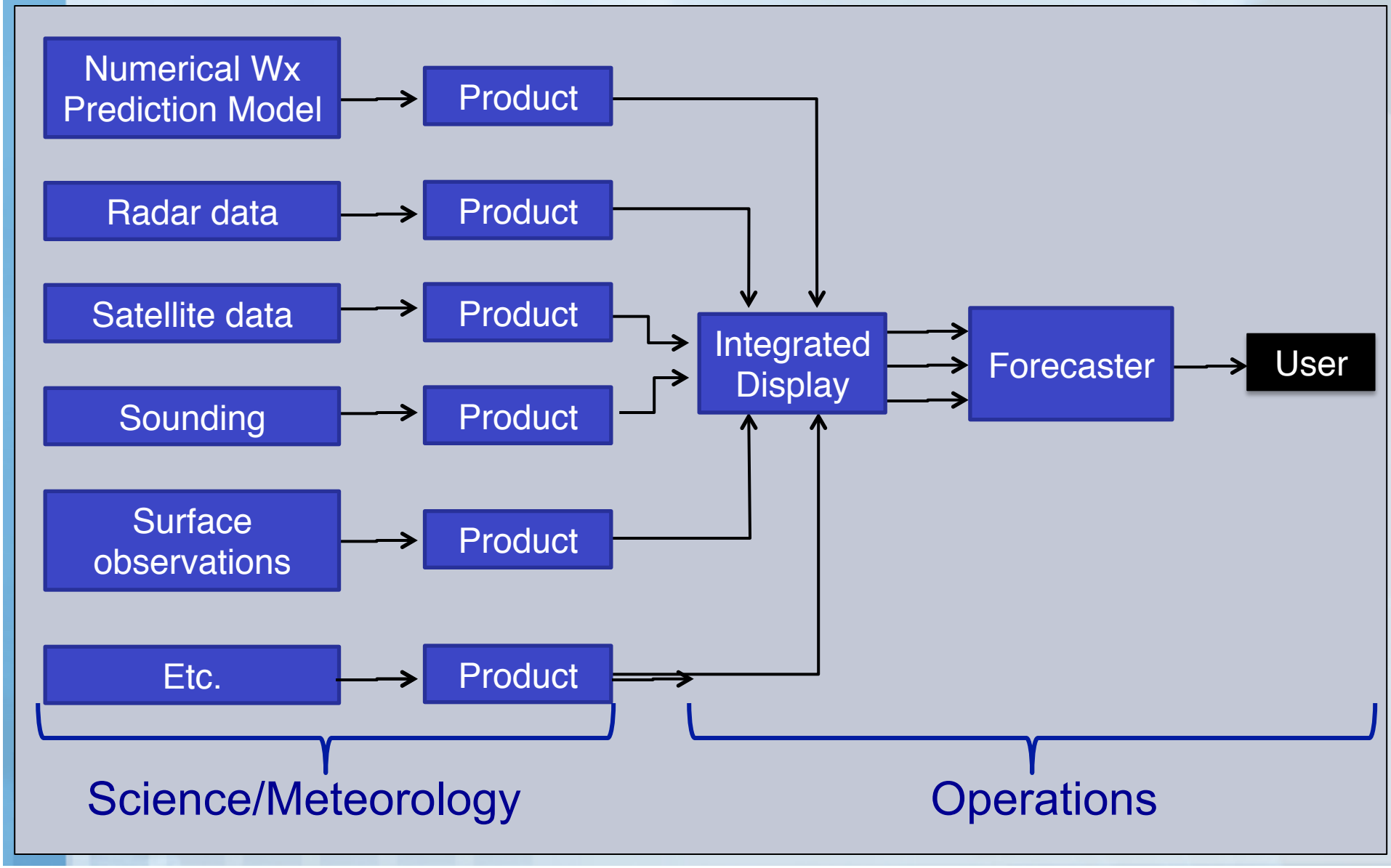
Weather disasters: Emergency managers

Weather forecasts: General public

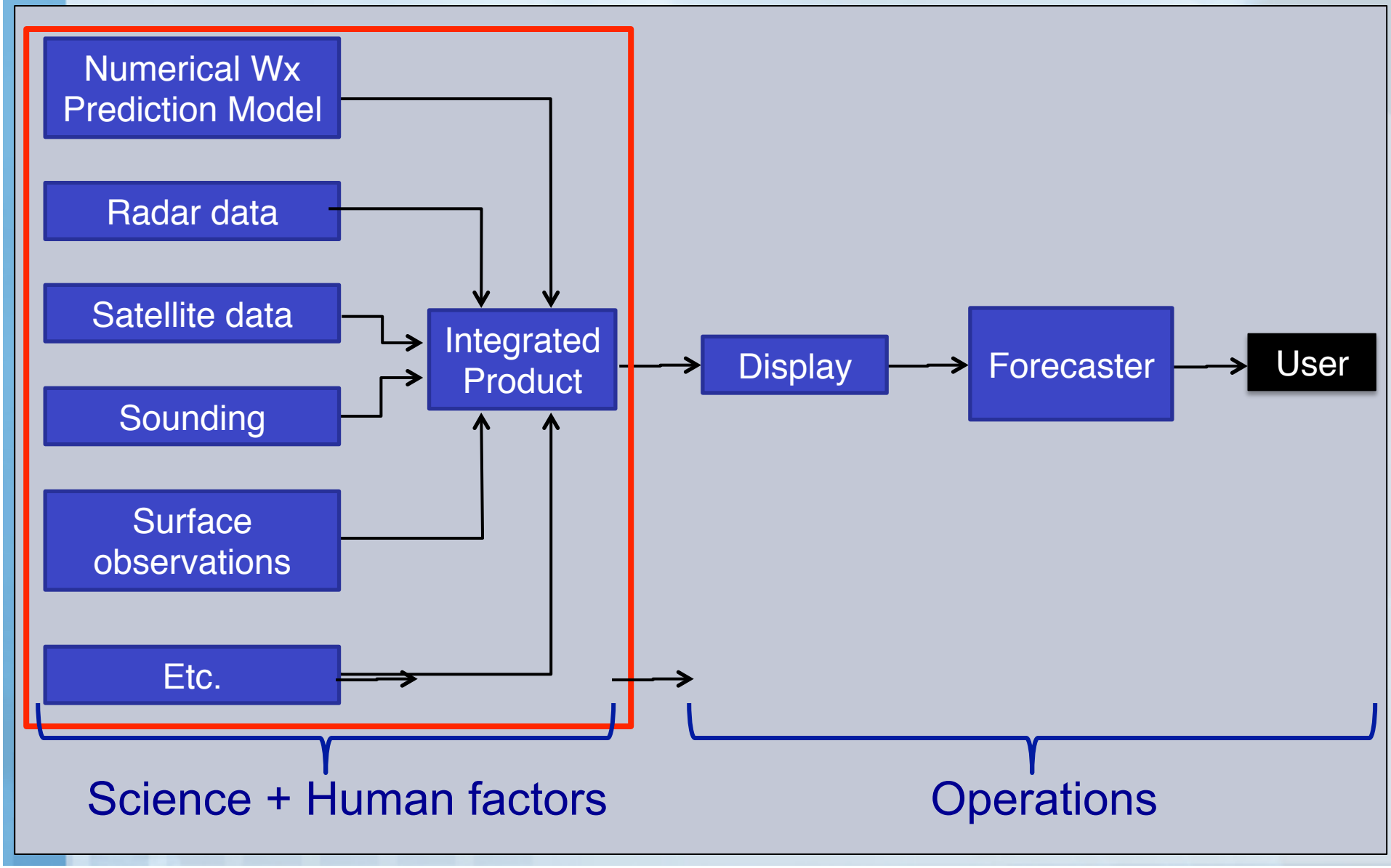
Real-time weather decision support: "Standard model"



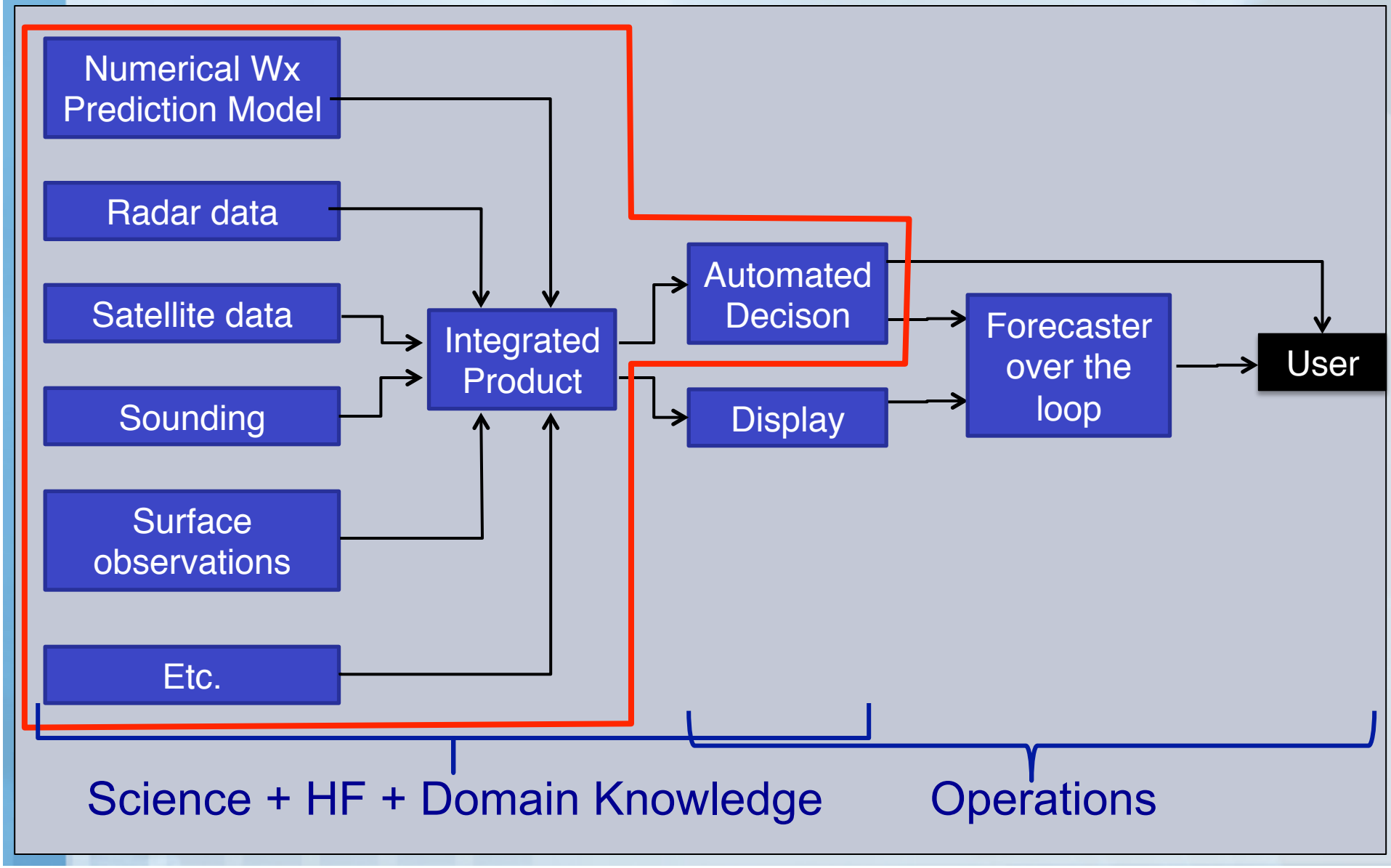
Real-time weather decision support:
"Enhanced standard model"

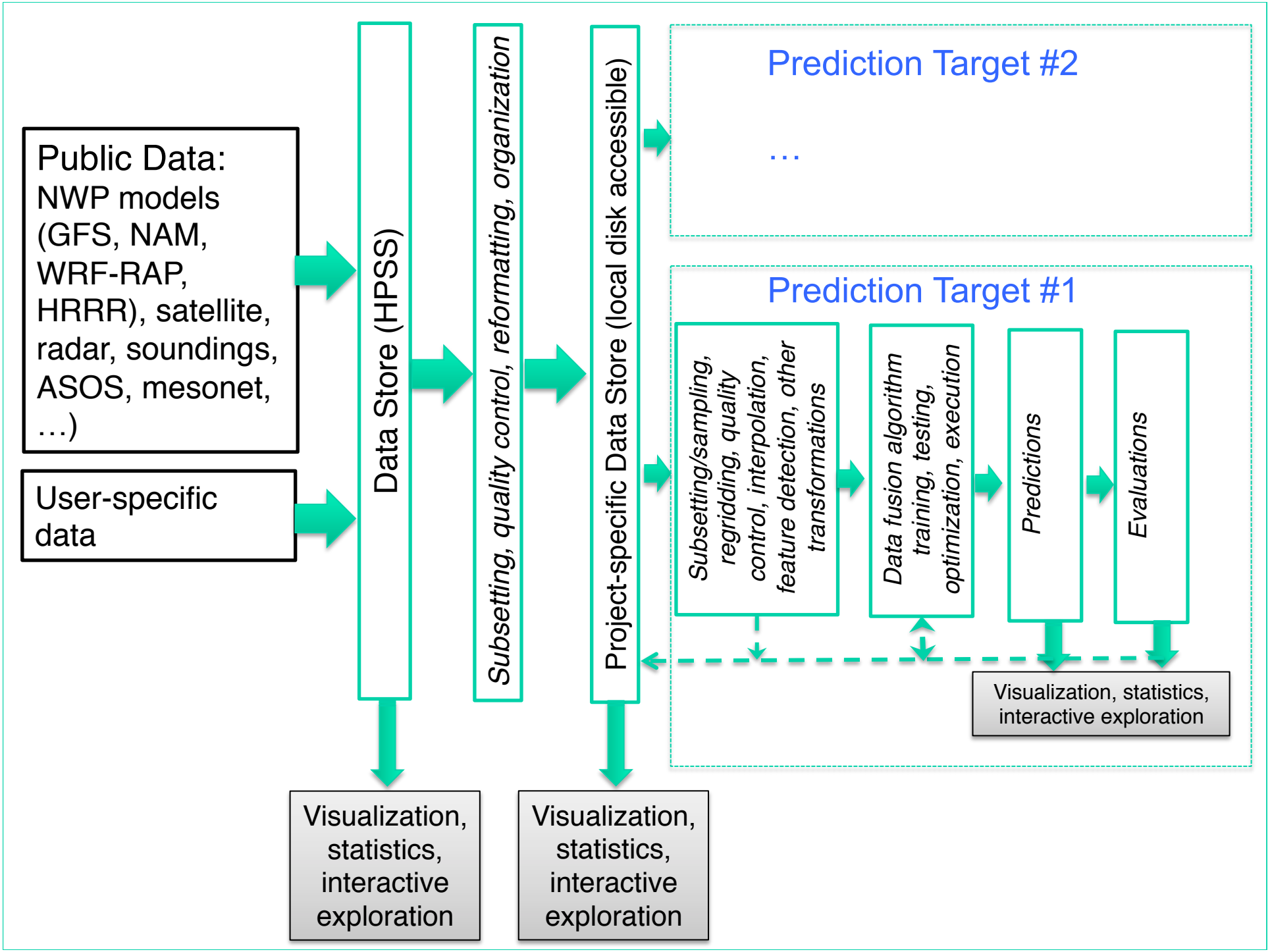


Real-time weather decision support:
Comprehensive integrated product



Real-time weather decision support:
Integrated product + automated decision

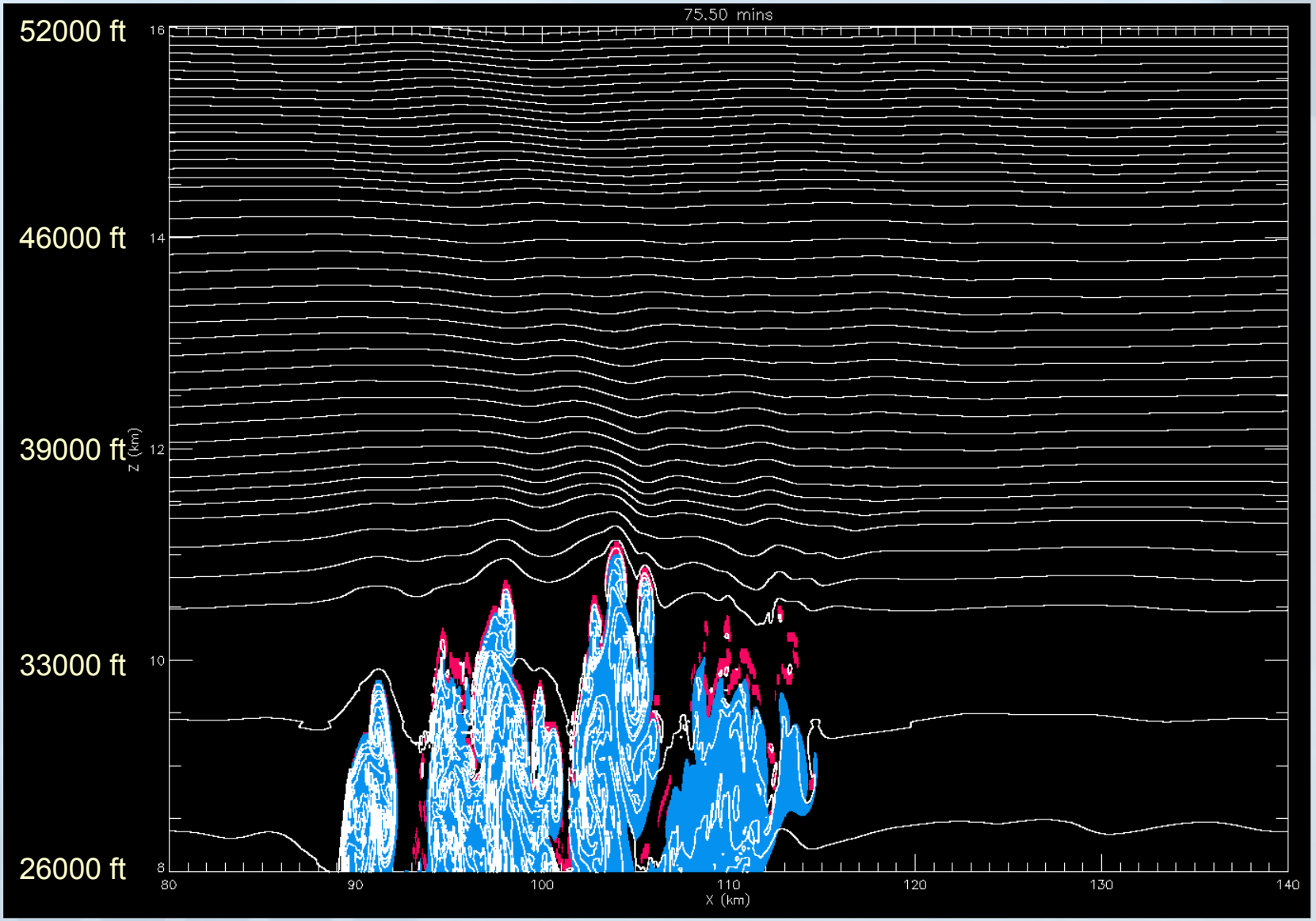



Public Data:
NWP models (GFS, NAM, WRF-RAP, HRRR), satellite, radar, soundings, ASOS, mesonet, …)
User-specific data
Data Store (HPSS)
Subsetting, quality control, reformatting, organization
Project-specific Data Store (local disk accessible)
Prediction Target #2
…
Prediction Target #1
Subsetting/sampling, regridding, quality control, interpolation, feature detection, other transformations
Data fusion algorithm training, testing, optimization, execution
Predictions
Evaluations
Visualization, statistics, interactive exploration
Visualization, statistics, interactive exploration
Visualization, statistics, interactive exploration

RF's Value to Empirical Modeling

- Identify *potential* contribution of variables
  - Independent of architecture of a custom algorithm
  - At different scales
  - In different scenarios (may identify "regimes")
- Confirm that an alternative algorithm is using available information efficiently through
  - Providing a performance benchmark
  - Evaluating variable importances in combination
- Create empirical predictive models
  - Appropriate to nonlinear problems involving complex, interacting sources of information
  - That can run reliably in real-time
  - That can be "easily" updated when inputs change (e.g., new NWP model, radar, satellite)

Use Case #1: Near-storm Turbulence



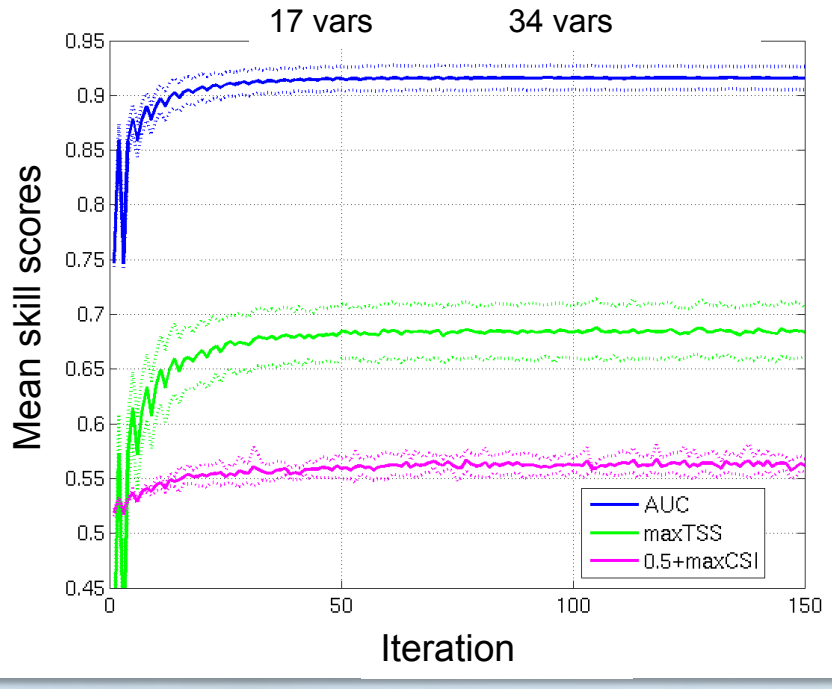
2-D simulation showing cloud, gravity waves, and turbulence (courtesy of Todd Lane)

Use Case #1: Near-storm Turbulence

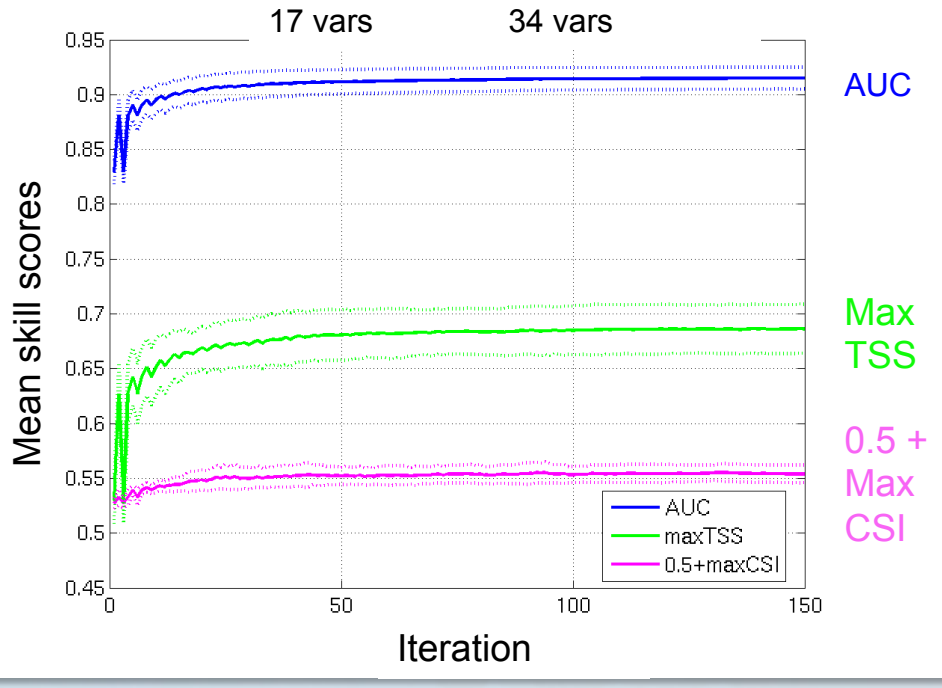
- Goal: develop Diagnosis of Convectively-Induced Turbulence (DCIT)
- Automated aircraft *in situ* EDR reports as "truth"
  - Millions per month, mostly null turbulence; resampled for modeling
- Fuse information from operational data sources:
  - near-storm environment fields and clear-air turbulence diagnostics derived from NWP models (e.g., WRF-RAP)
  - lightning and satellite-derived features and turbulence signatures (e.g., overshooting tops)
  - storm features from Doppler radar in-cloud turbulence (NTDA) and 3-D radar reflectivity
  - distances to contours, disc neighborhood statistics
  - Use MySQL DB to collect collocated data (hundreds of GB)
- Use statistical learning methods to select predictors, build and test an empirical model for turbulence diagnosis

Variable Selection

- Used RF "importance" rankings to identify top 107 of 1200+ predictor variables for predicting EDR ≥ 0.2 $m^{2/3}$ s^{-1} (moderate)
- Sampled training/testing subsets from odd/even Julian days
- Performed variable selection (iterated 2 forward selection steps, 1 back) for both 50-tree random forest and logistic regression
- Repeated for 8 training/testing samples; aggregated results



Random forest variable selection



Logistic regression variable selection

Identifying Top Predictor Variables

Random forest top predictors

Rank	Mean Occ.	Predictor variable name
1	115	Dist. to NSSL echo top >10 kft
2	114	Model FRNTGTHRI
3	104	Model RITW
4	89	Model ELLROD2
5	88	Diff. Alt. to 80-km max NTDA sev. top
6	85	Model MWT2
7	79	Model ELLROD1
8	78	160-km mean of Satellite Ch. 6
9	69	Model F2DTW
10	68	Model MWT3
11	68	40-km min of Satellite Ch. 6
12	67	Model Atm. Pressure
13	66	Model BROWN2
14	65	Satellite Ch. 4 minus Model temp.
15	64	Model DUTTON
16	63	Satellite Ch. 4 minus Satellite Ch. 3
17	58	Model NGM2
18	56	160-km mean of Satellite Ch. 4
19	53	Model RICH
20	52	Diff. Model pres. to Mod. surf. pres.

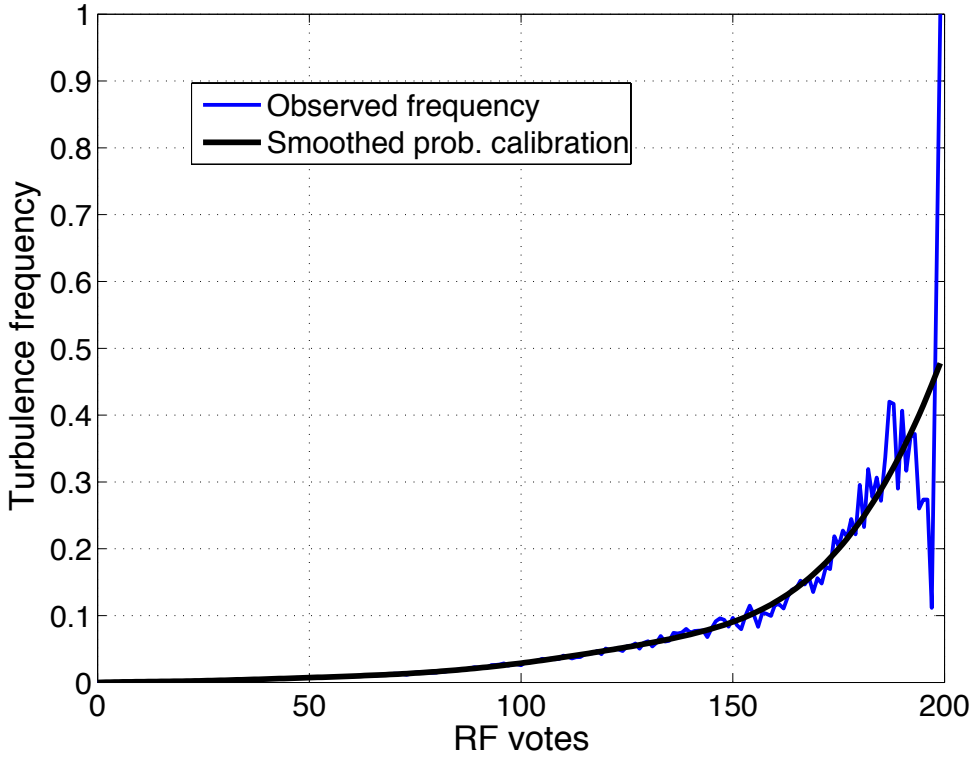
Logistic regression top predictors

Rank	Mean Occ.	Predictor variable name
1	135	Model FRNTGTHRI
2	134	Diff. Alt. to 80-km max NTDA sev. top
3	127	Dist. to echo top >10 kft
4	126	10-km max of NSSL echo top
5	121	Model ELLROD2
6	111	Model RITW
7	107	Model BROWN2
8	94	Diff. Alt. to 20-km max NTDA mod. top
9	92	Model BROWN1
10	89	Model ELLROD1
11	88	Model MWT3
12	87	Model EDRRI
13	85	Model DTF3
14	83	20-km no. of good NTDA dBZ points
15	77	10-km no. of good NTDA dBZ points
16	76	10-km mean of NTDA composite EDR
17	74	10-km max of NTDA composite EDR
18	74	10-km min of Satellite Ch. 3
19	73	10-km mean of NSSL echo top
20	69	Model IAWINDRI

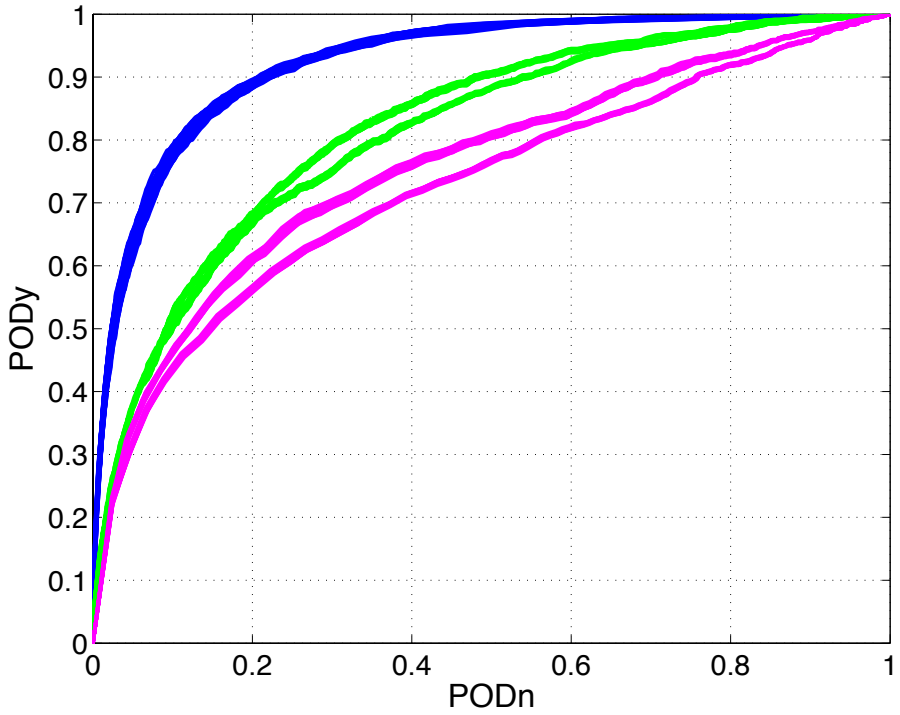
- Top predictor sets are different: random forest results include fields not monotonically related to turbulence (e.g., pressure)

Calibration and Evaluation

- Based on 32 cross-evaluations using even/odd and odd/even Julian day training/testing sets



RF votes to probability calibration



ROC curves for RF DCIT (blue), GTG (green), and storm distance (magenta)

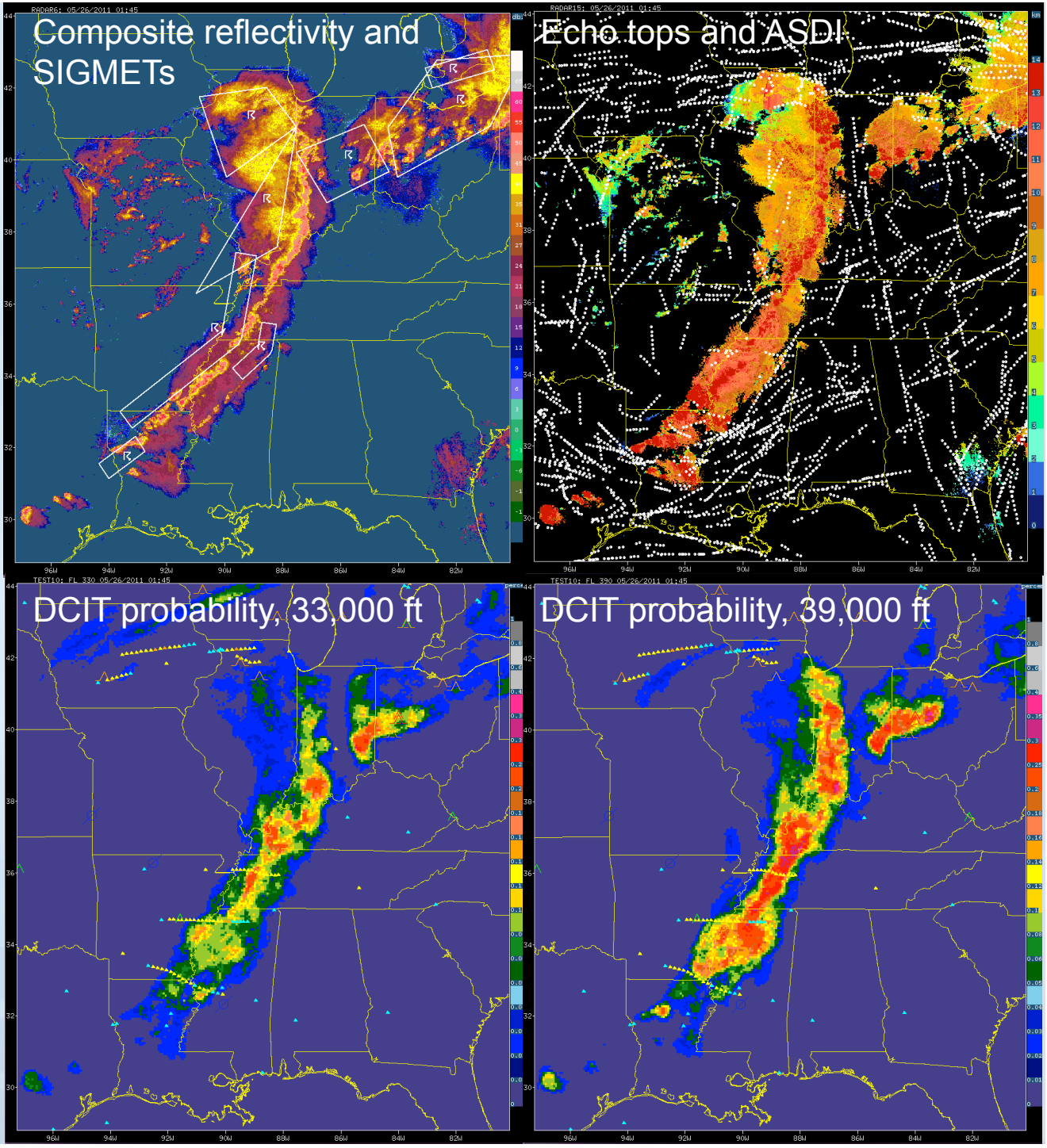
Evaluating Performance

- Compared 200-tree RF with NWP model-based Graphical Turbulence Guidance v. 3, KNN, logistic regression and distances to "storms" (regions with echo tops > 10,000 ft)
- Evaluation performed on 32 cross-evaluation subset pairs from odd and even Julian days

DAL upper level skill scores (32)						
Method	**AUC**	**Std**	**MaxCSI**	**Std**	**MaxTSS**	**Std**
RF	0.924	0.002	0.075	0.006	0.699	0.006
KNN	0.915	0.001	0.064	0.003	0.688	0.003
LR	0.915	0.004	0.060	0.005	0.677	0.011
GTG 3	0.816	0.008	0.034	0.002	0.483	0.011
Storm distance	0.743	0.016	0.021	0.0005	0.389	0.026
DAL lower level skill scores (32)						
Method	**AUC**	**Std**	**MaxCSI**	**Std**	**MaxTSS**	**Std**
RF	0.911	0.002	0.137	0.005	0.678	0.004
KNN	0.895	0.002	0.114	0.005	0.645	0.004
LR	0.893	0.002	0.112	0.003	0.638	0.004
GTG 3	0.736	0.005	0.069	0.002	0.377	0.011
Storm distance	0.650	0.009	0.042	0.003	0.243	0.009

DCIT Case Study

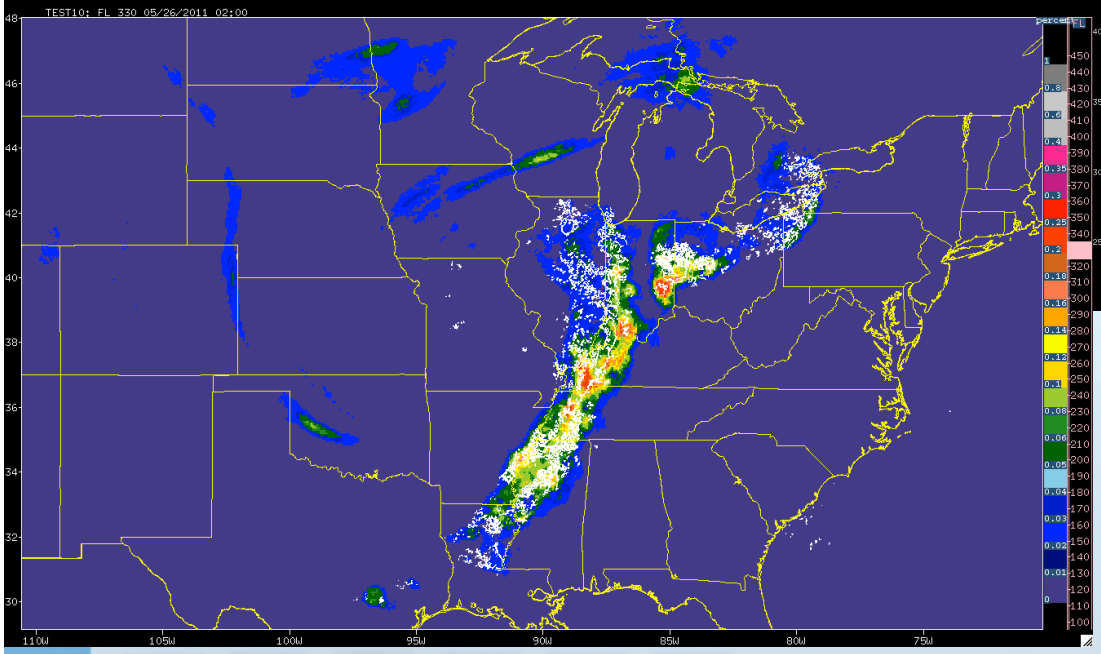
- May 26, 2011 01:45 UTC



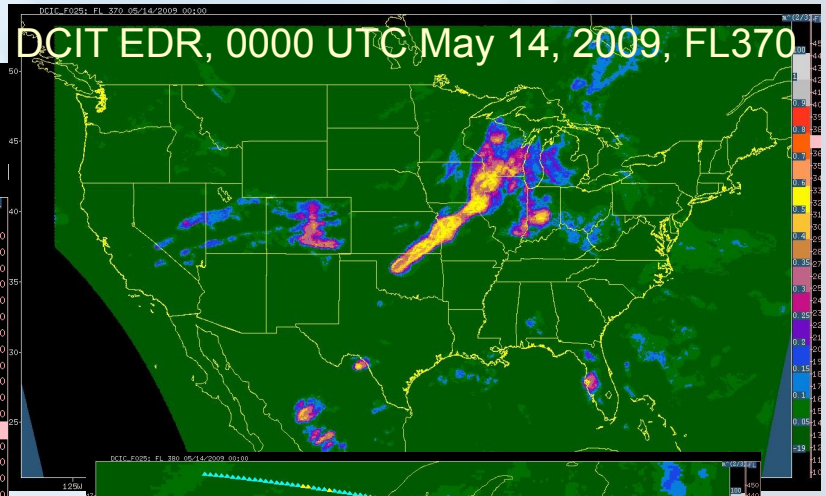
Real-time DCIT

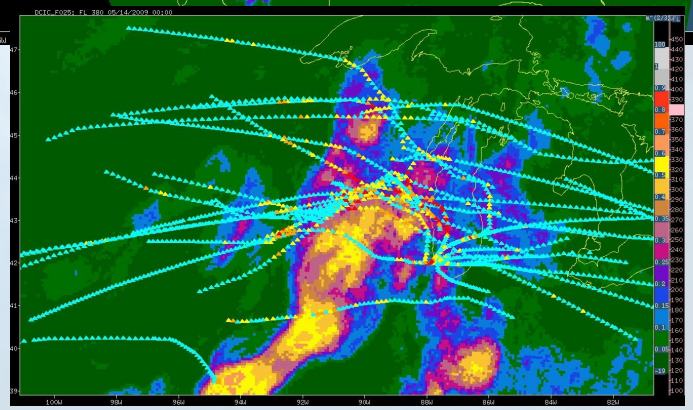
- Runs at NCAR/RAL in successive versions since 2008
  - 15-min update, 6 km horizontal, 1,000 ft vertical resolution
  - calibrated to deterministic EDR for GTG Nowcast
  - designed to use different statistical models for different altitudes, data availability

DCIT prob. of light to moderate or greater turb.
0200 UTC May 26, 2011, FL330 (w/10 kft echo top contours)



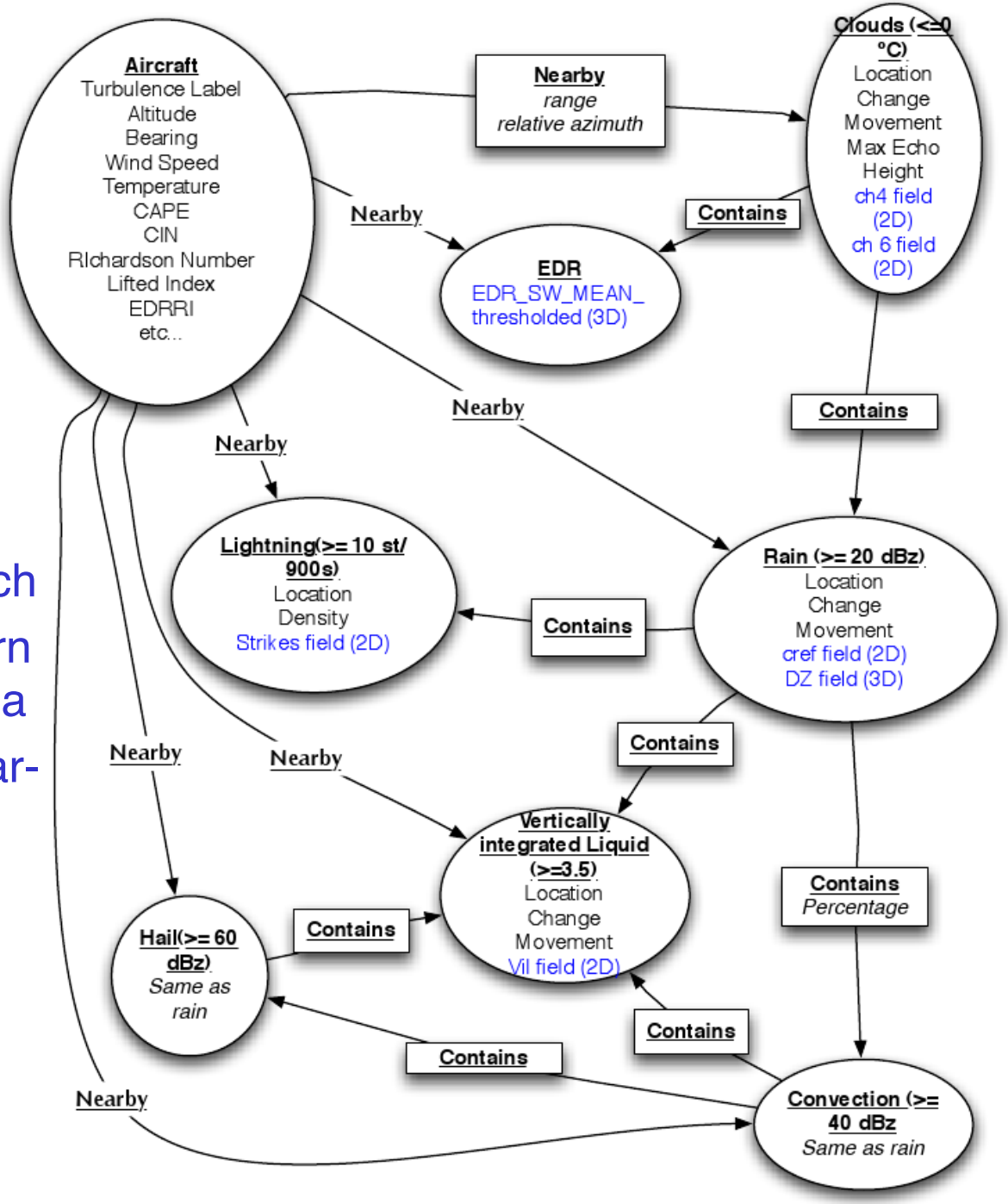
DCIT EDR, 0000 UTC May 14, 2009, FL370





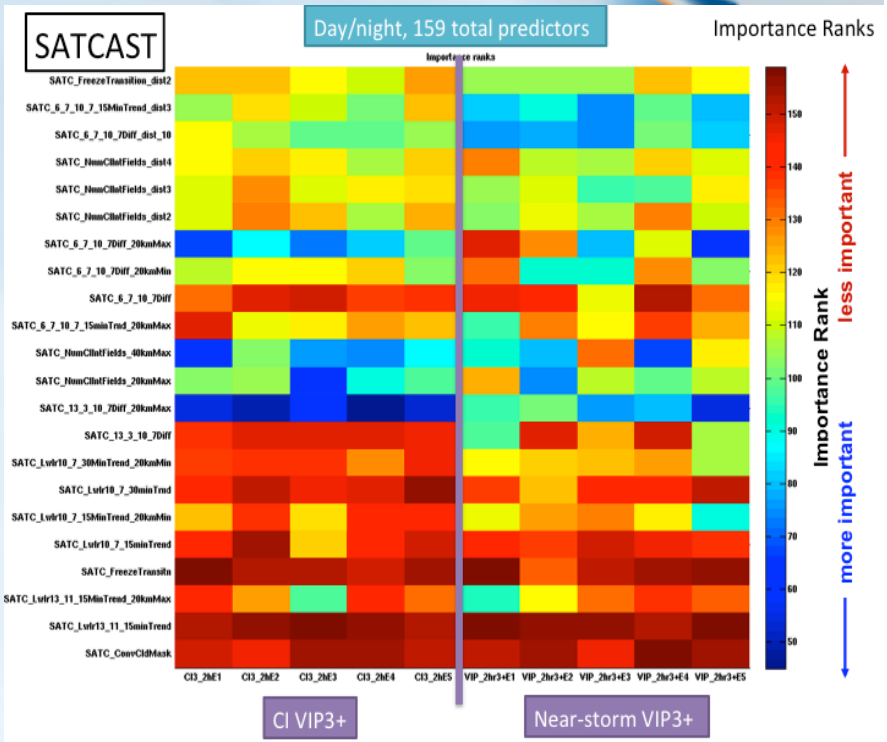
Spatio-temporal Relational Probability Trees (SRPTs) and Random Forests (SRRFs)

- Object relational approach
- Work with Amy McGovern at University of Oklahoma
- Example schema for near-storm turbulence prediction
- Future extension: vary definitions of objects themselves to more fully automate exploration



Use Case #2: Nowcasting Storms

- Truth: radar vertically-integrated liquid (VIL) advected backwards to be coincident with antecedent observations
- Predictor variables: observations, NWP models and derived fields and features
- Many TB of truth and predictor data
- Resample and use random forest methodology to predict VIL level at each pixel at 1 and 2 h lead times
- Use importances to choose regimes, variables
- Outperforms competitive forecasts
- *Weakness:*
  - *Difficulty of aligning predictors to predict isolated, rare events like convective initiation*

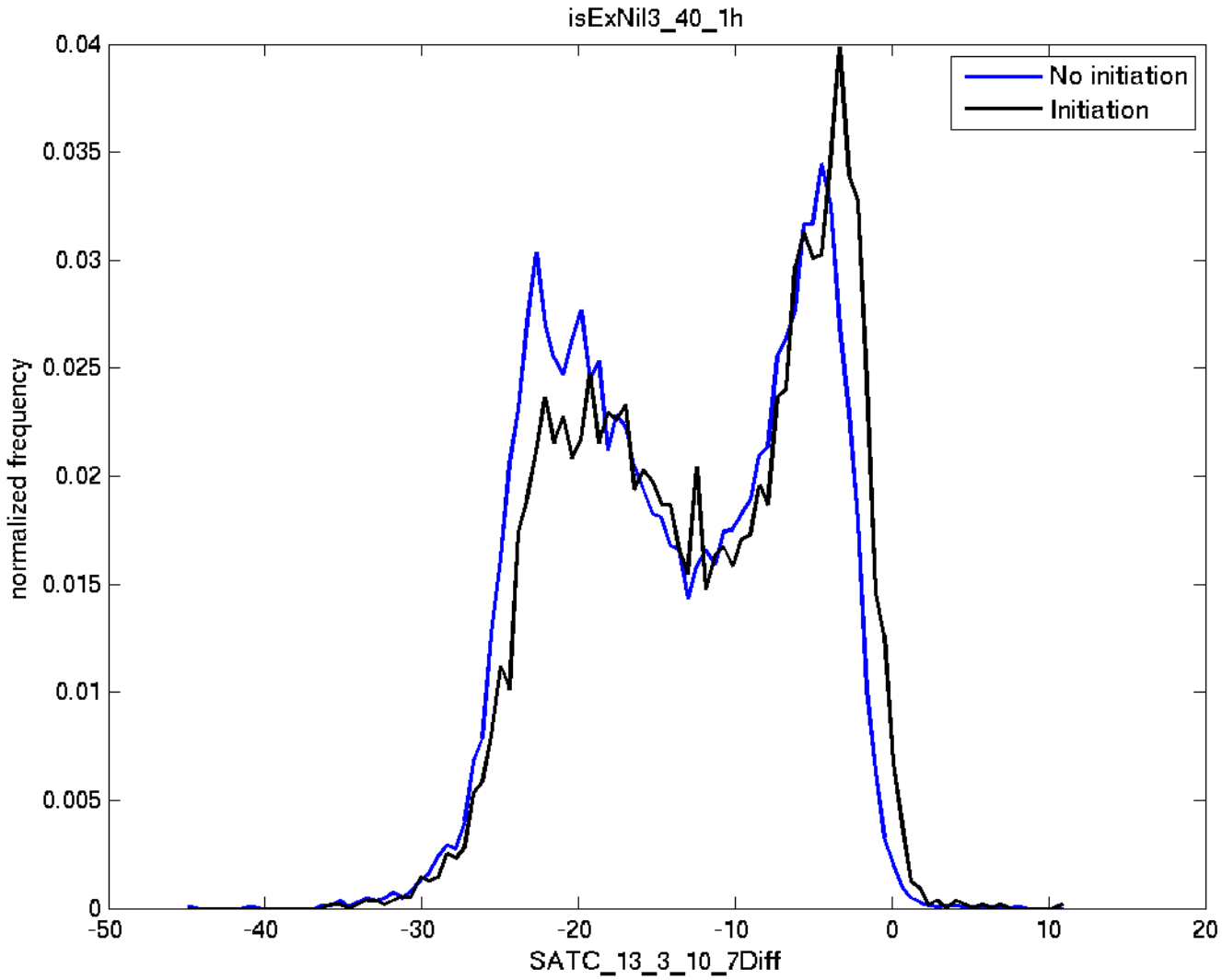


Predicting Convective Initiation	Max CSI	Max TSS	AUC
2h simple extrapolation	0.005 ± 0.002	0.17 ± 0.05	0.60 ± 0.03
CoSPA (2h)	0.012 ± 0.005	0.12 ± 0.03	0.56 ± 0.02
LAMP 1-3h (2hr)	0.023 ± 0.006	0.56 ± 0.03	0.83 ± 0.01
2h RF	**0.032** ± 0.011	**0.68** ± 0.02	**0.91** ± 0.01

CSI = Critical Success Index, TSS = True Skill Score, AUC = Area Under the ROC Curve

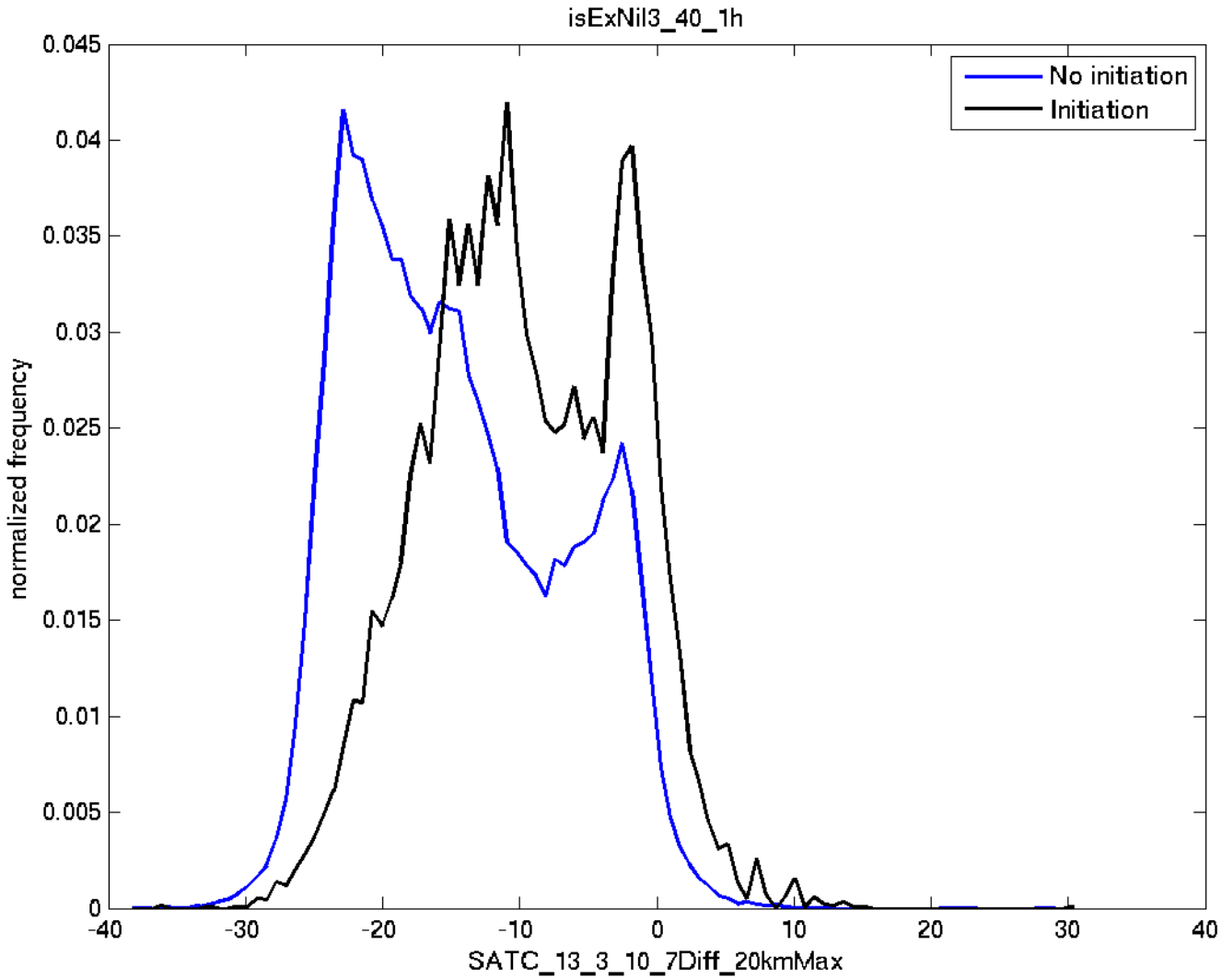
Conditional Histograms

13.3-10.7 micron, 1-hr, 40 km VIP 3 Ex Nihilo initiation

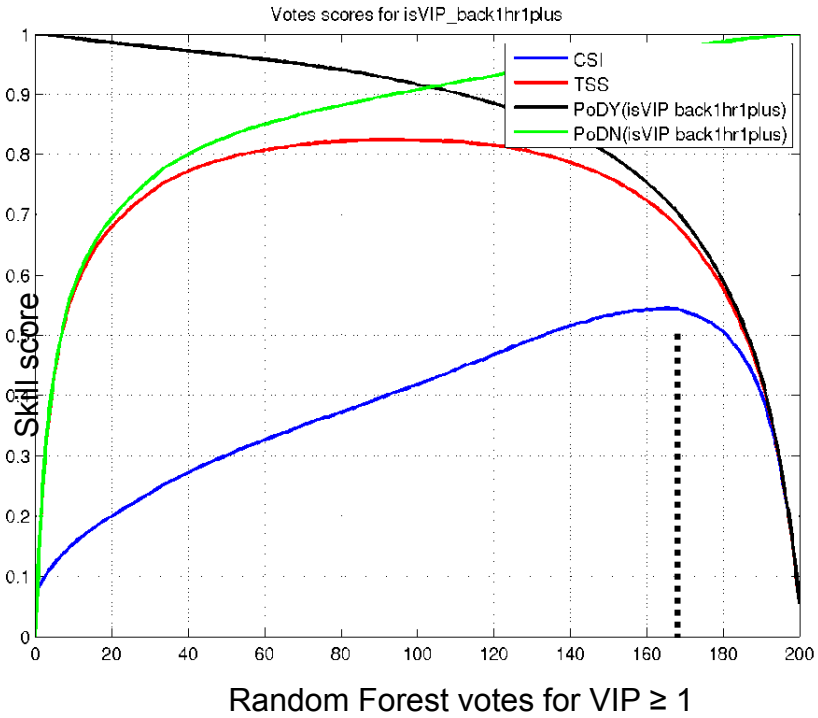


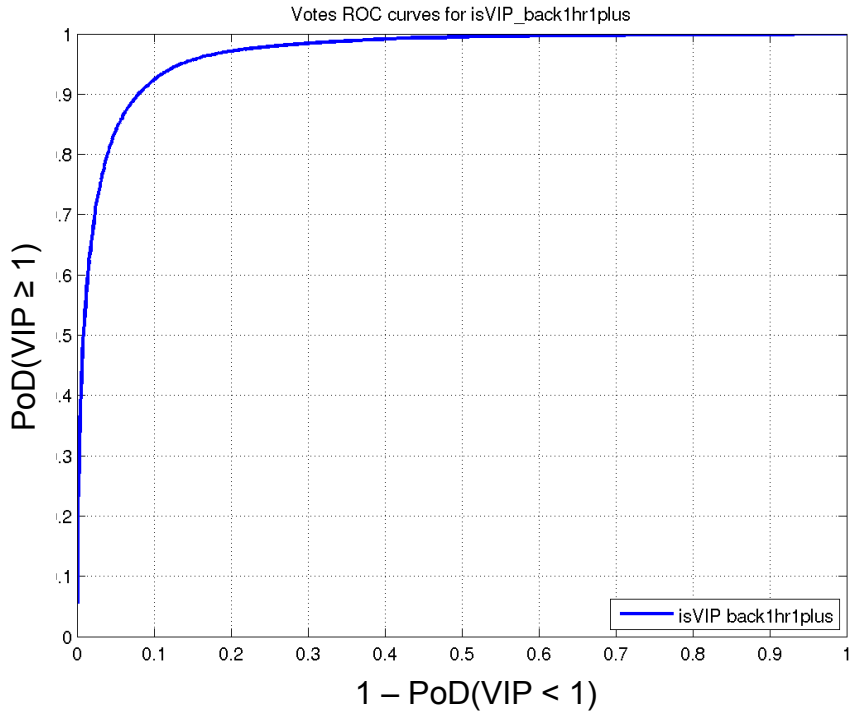
Conditional Histograms

13.3-10.7 micron 20 km Max, 1-hr, 40 km VIP 3+ Ex Nihilo initiation



RF Performance: VIP 1+



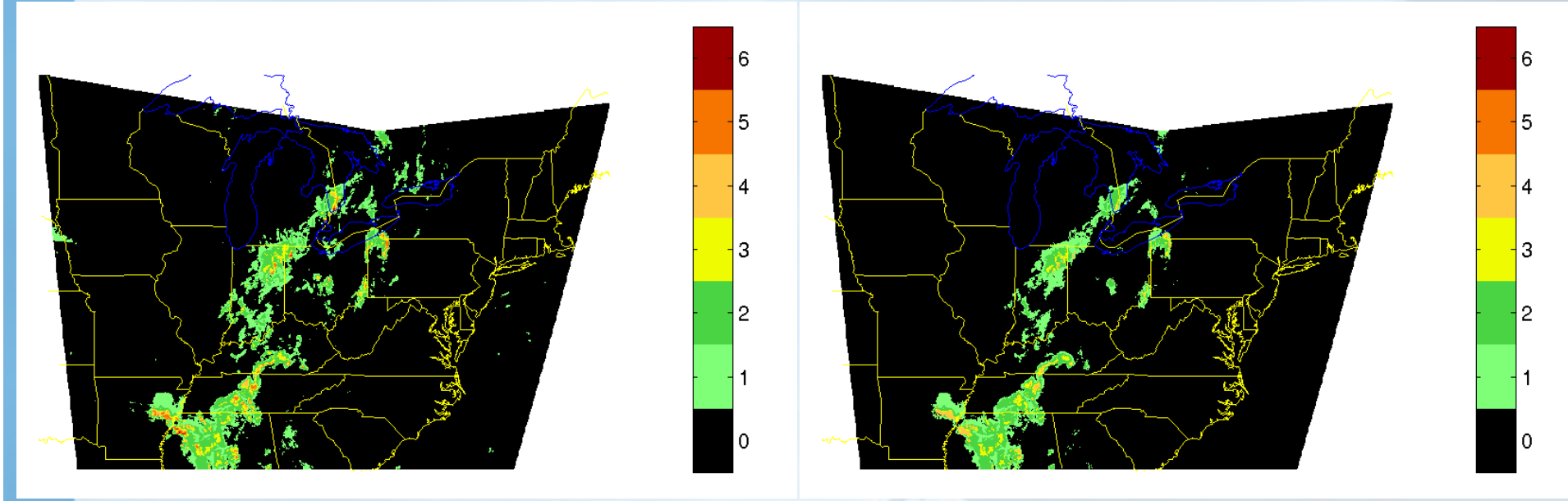


Skill scores as a function of number-of-votes threshold

ROC curve, created by varying number-of-votes threshold

- Use maximum CSI threshold (dotted line on left plot) to create deterministic prediction contour

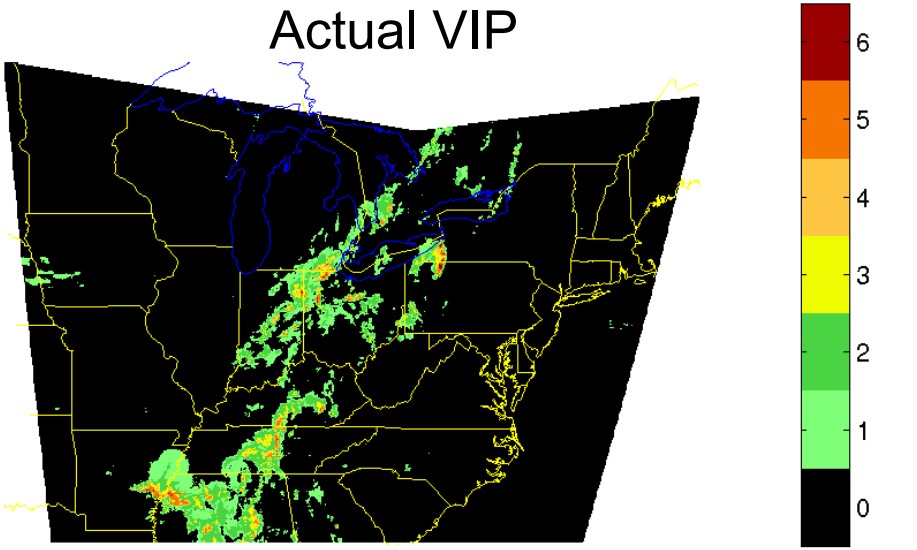
Case: 20070619, 1405 UTC, VIP



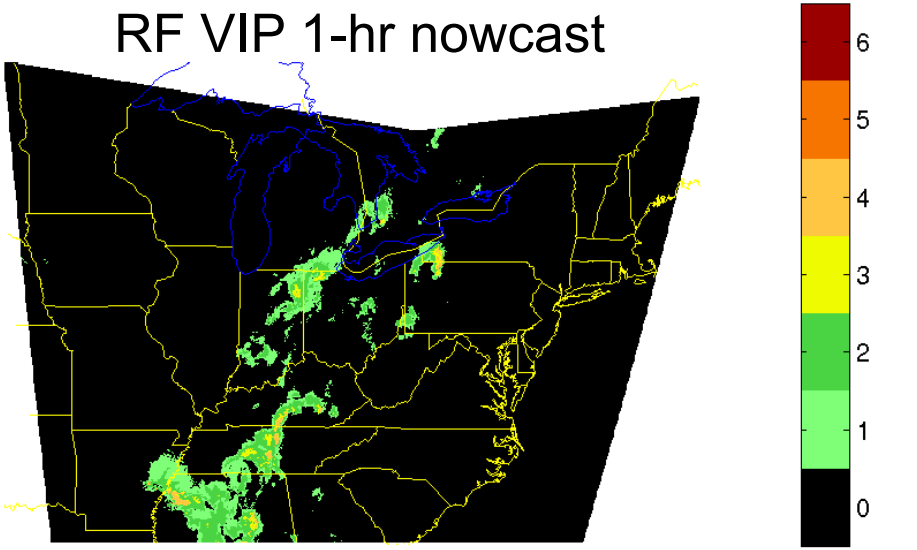
- For each RF VIP threshold-exceedance nowcast (1+, 2+, 3+, 4+, 5+, 6+), choose threshold that maximizes CSI skill score to make a binary forecast
- Create a deterministic VIP level composite nowcast

Case: 20070619, 1505 UTC, VIP

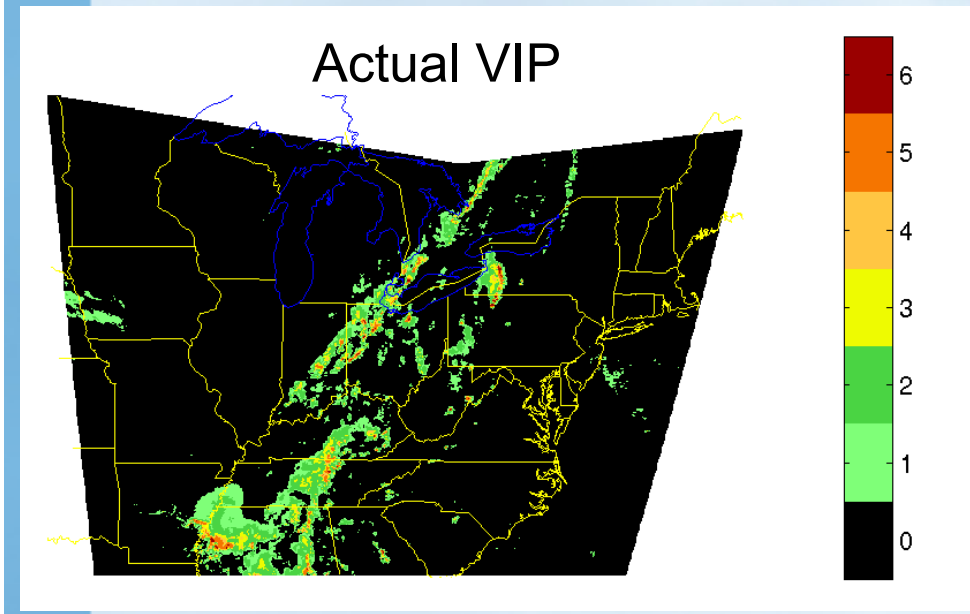
Actual VIP

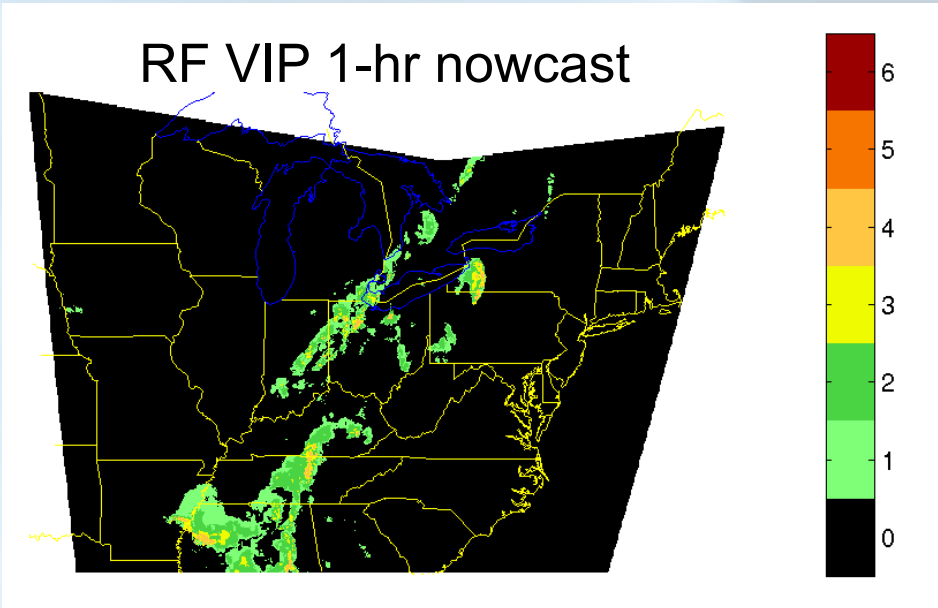


RF VIP 1-hr nowcast

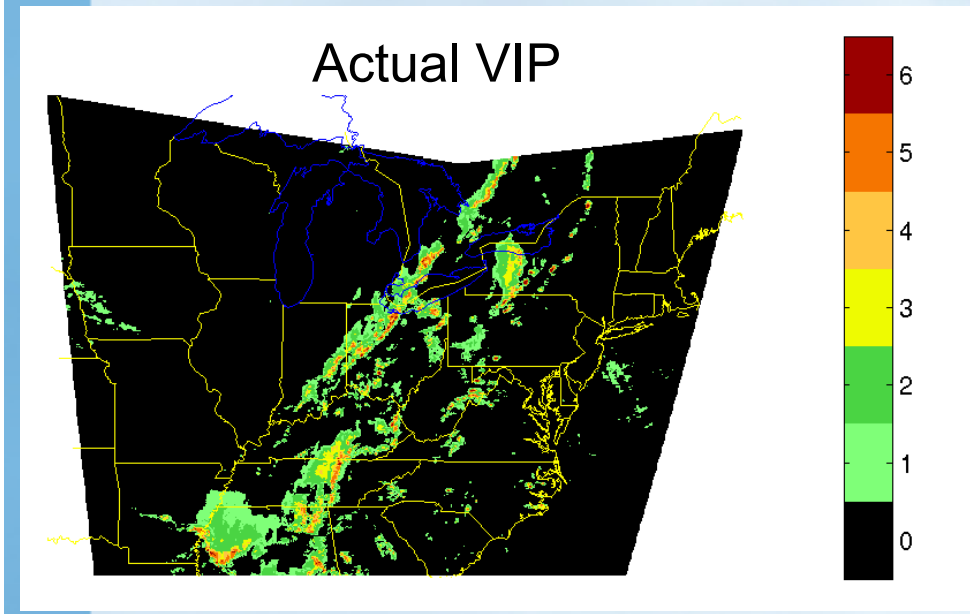


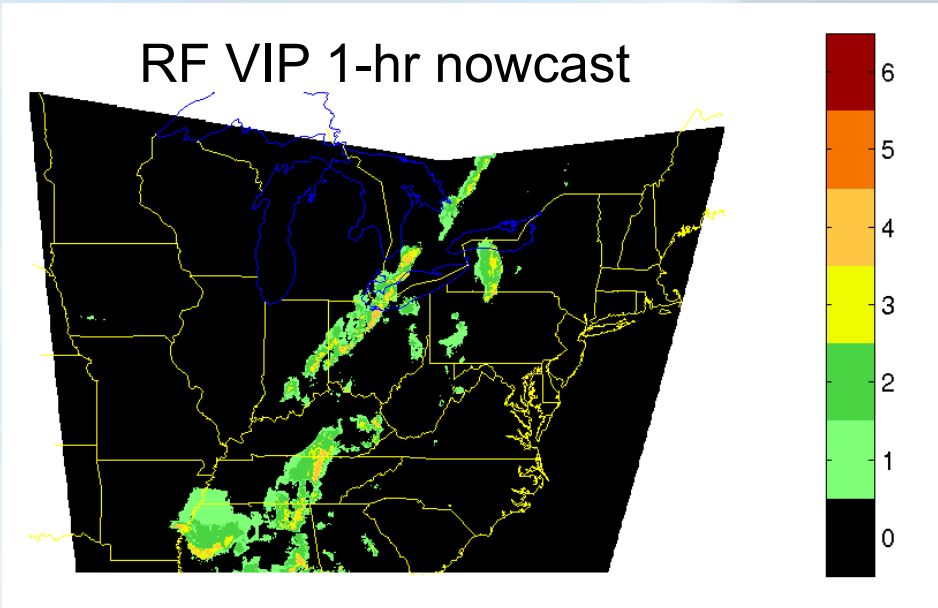
Case: 20070619, 1605 UTC, VIP



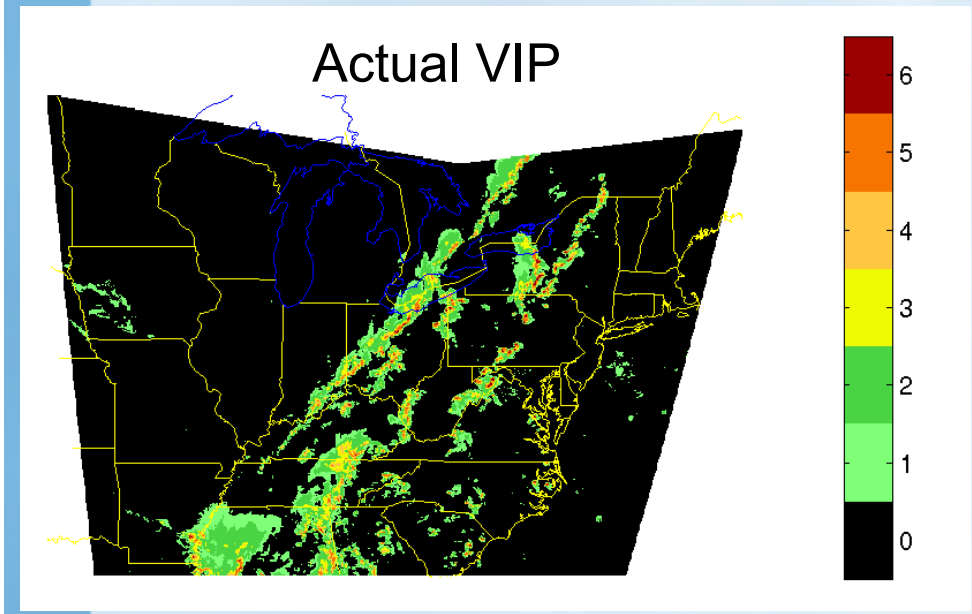


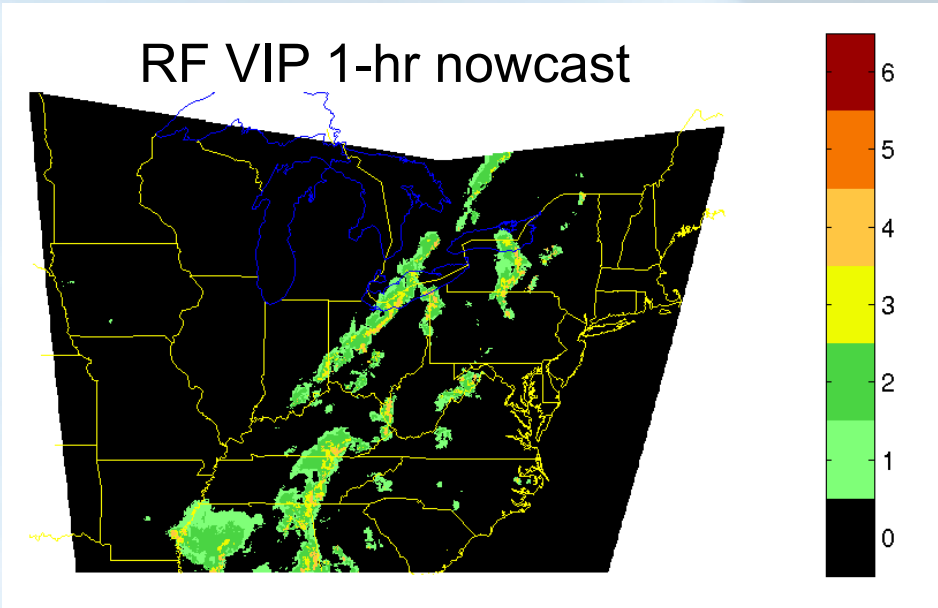
Case: 20070619, 1705 UTC, VIP



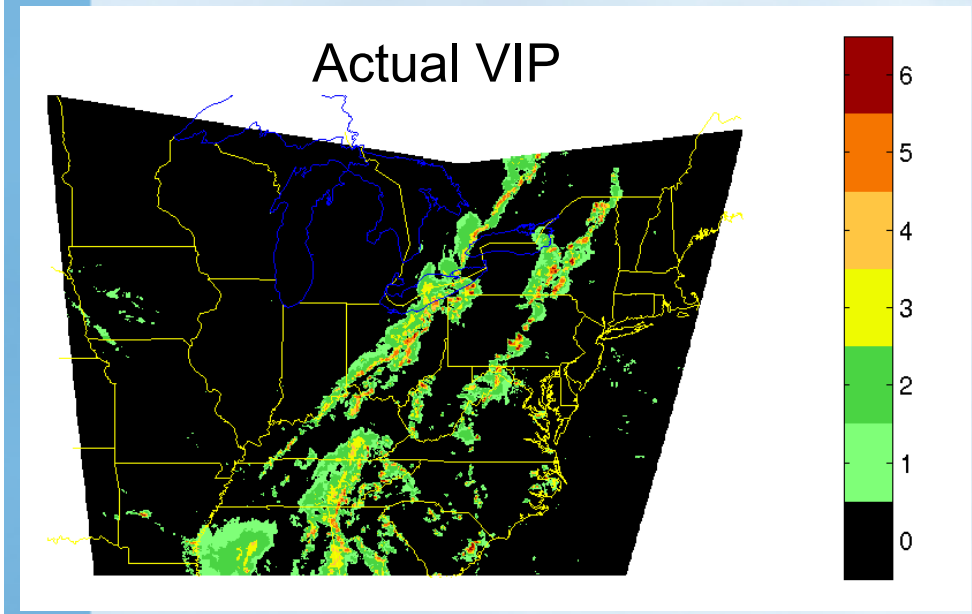


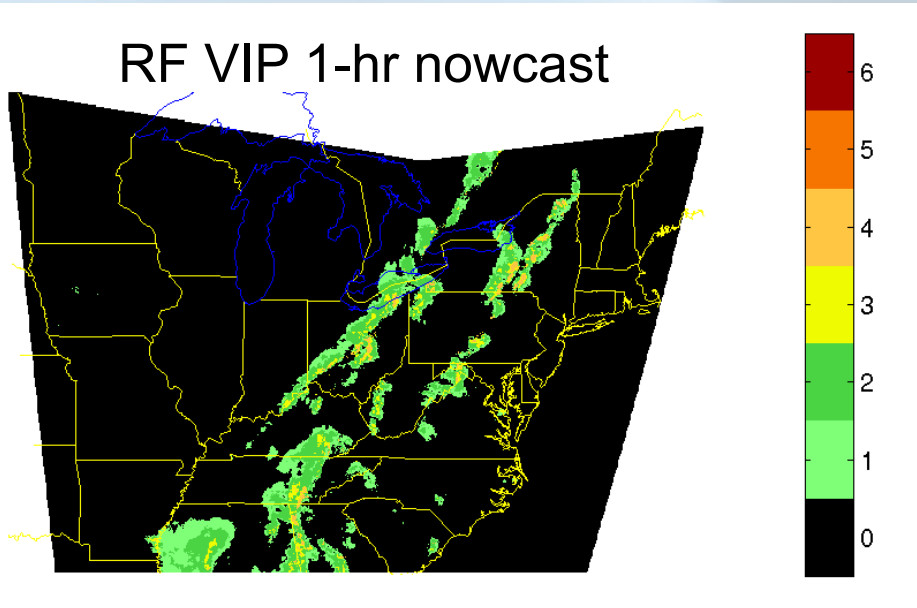
Case: 20070619, 1805 UTC, VIP





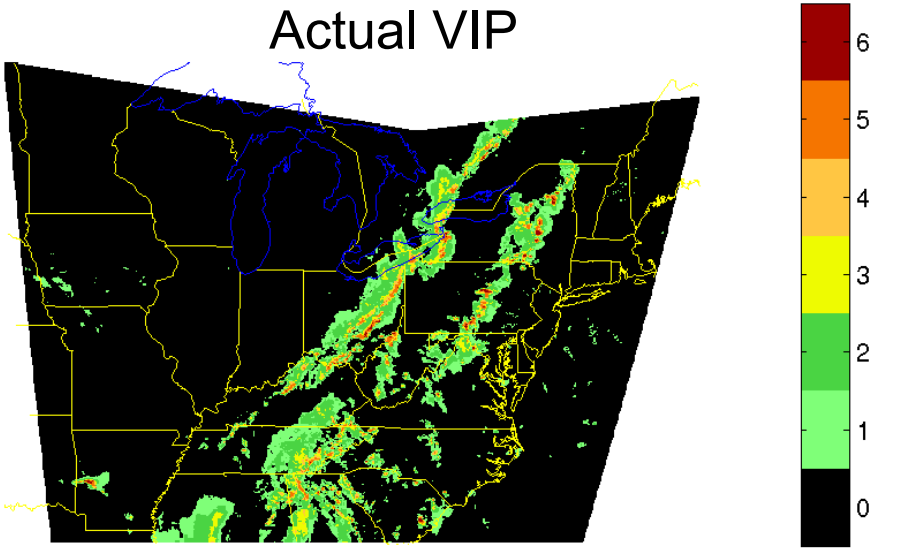
Case: 20070619, 1905 UTC, VIP



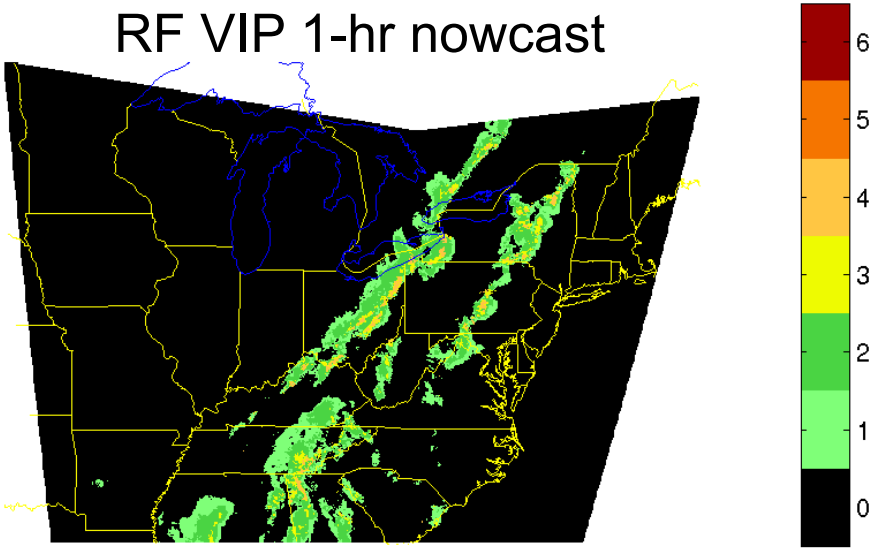


Case: 20070619, 2005 UTC, VIP

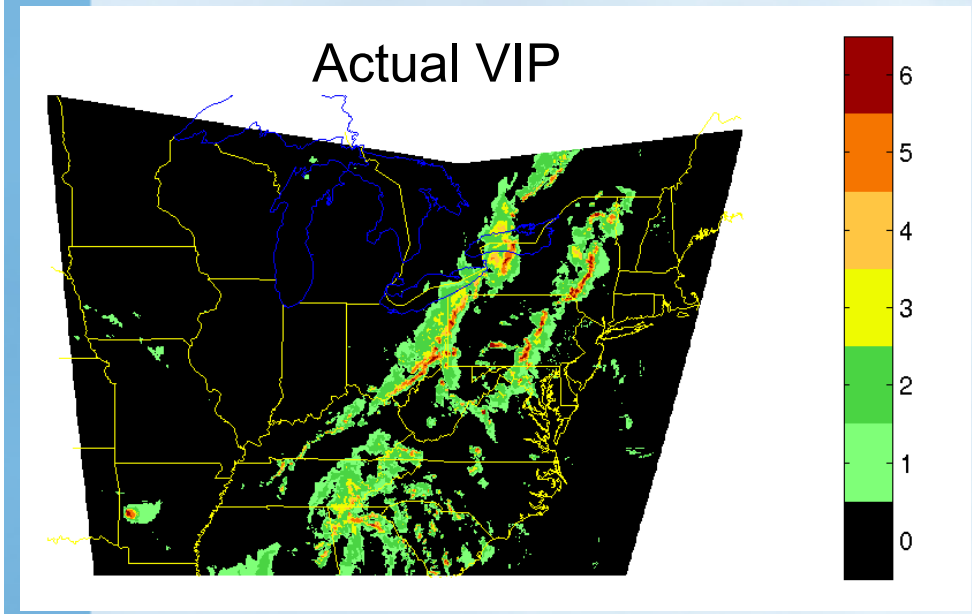
Actual VIP

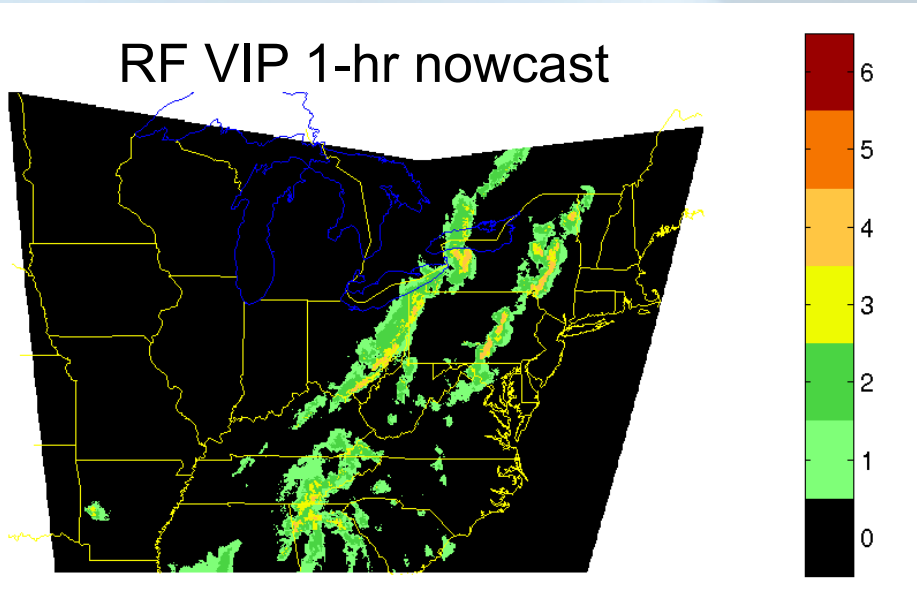


RF VIP 1-hr nowcast

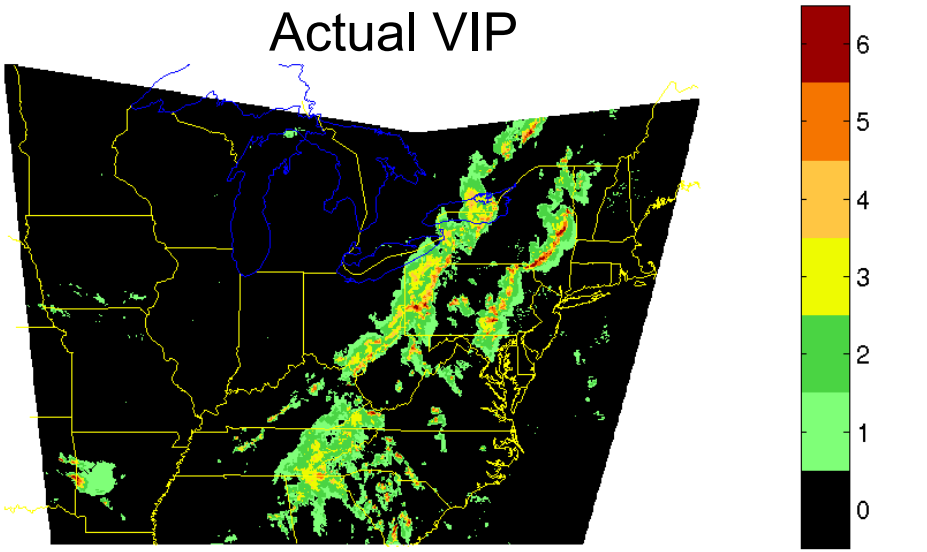


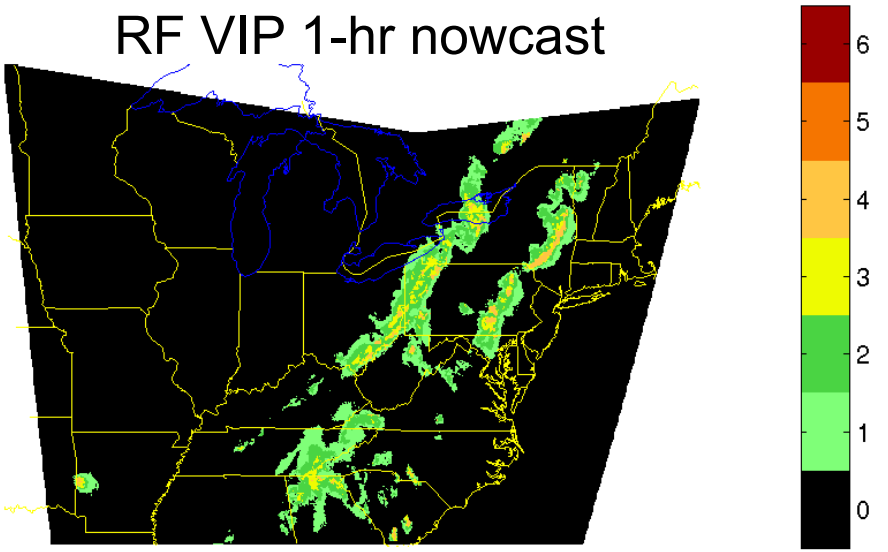
Case: 20070619, 2105 UTC, VIP





Case: 20070619, 2205 UTC, VIP





Use Case Summary

- Human domain knowledge is used to compute relevant features from raw data fields
- Statistical learning is used to identify a minimal set of predictors and create an empirical model optimized relative to a user-relevant skill score
- Undersampling of common events is used to ensure adequate discrimination of rare events by a statistical classification algorithm
  - Calibration to generate desired output quantity
- Real-time system processes and aligns data on arrival, chooses appropriate predictive model based on regime and data availability

SE Challenges for Big Data Stat. Learning

NCAR

- Reliable access to disparate sources of high-quality data with potential predictive value
  - Adequate organization, metadata, accessibility suitable for automated discovery and exploration
  - Server-side subsetting or summarizing of data
- Efficient methods for asynchronous, inconsistent data
  - Data quality control
  - Deriving and tracking relevant features in space and time
  - Computing time-evolution rates and trends, including in 3D
  - Alignment of various observation and model data
    - E.g., adjust for movement over time lags, satellite parallax, satellite pixel size varying with latitude, model/obs phase errors, ensembles
    - Avoiding interpolation “smearing” to the extent possible
- Flexible server-side computational capabilities

SE Challenges for Big Data Stat. Learning

NCAR

- Scale or develop methods and workflows for discovering patterns and predictive relationships in 3-D + time data
  - E.g., would like to discover relevant objects, attributes and relations to predict evolution of storms and associated hazards: e.g., lightning, airframe icing, and turbulence
  - Optimize user-relevant metrics, not just RMSE, etc.
    - E.g., object-oriented metrics (MET-MODE), economic models
  - Handle rare (extreme) events and risk assessment
  - Improved DB for storing/organizing intermediate fields
- Enhanced methods for extracting "culturally-relevant" scientific knowledge
  - Simple or "conceptual" models based on relevant features
  - Human-interactive discovery process
- Fast, robust, autonomous systems for real-time online empirical model development and execution

Resources and Opportunities

- Random forest implementations
  - PARF (code.google.com/p/parf/)
  - Original Fortran (http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm)
  - WEKA (www.cs.waikato.ac.nz/ml/**weka**/)
  - R (randomForest, cforest)
  - MATLAB ("treebagger")
- NSF EarthCube: http://earthcube.ning.com
- Climate Informatics Workshop: www2.image.ucar.edu/event/ci2013
- AMS AI Committee: http://ai.metr.ou.edu
- Nascent RAL Big Data Analysis and Statistical Learning Group: contact me (jkwillia@ucar.edu)

Acknowledgement

- This research is supported in part by NASA Applied Sciences. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.
- This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

Contact information

John K. Williams

National Center for Atmospheric Research

jkwillia@ucar.edu

303-497-2822