Emerging Methods for PostProcessing

John K. Williams
The Weather Company/WSI

Workshop on The Future of Statistical Postprocessing in NOAA and the Weather Enterprise

College Park, MD
January 20, 2016

Why do post-processing?

- Create more accurate deterministic forecasts
  - Remove systematic error
    - Magnitude, location, timing
  - Combine information from multiple models
  - Downscale to user-relevant times and location
- Supply uncertainty information
  - Error bars, PDFs, set of exemplars
- Predict derivative quantities
  - Rare, high-impact, or multivariate events
  - Decision support products
- “Good forecast” depends on the application

Consensus Weather Forecasting

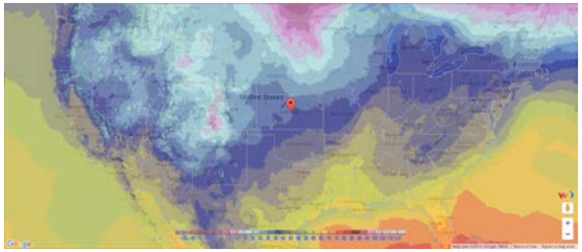
- Use ensemble of NWP and MOS forecasts to construct more accurate predictions
  - bias adjust to account for systematic error
  - average using weights based on input forecast accuracy and independence

$$\overbrace{F}^{\text{Consensus forecast}} = \sum_{i=1}^{N} \underbrace{\Big(\overbrace{f(i)}^{\text{Input forecast}} - \overbrace{b(i)}^{\text{Bias estimate}}\Big)}_{\text{Bias-corrected input forecast}} \overbrace{w(i)}^{\text{Weight}}$$

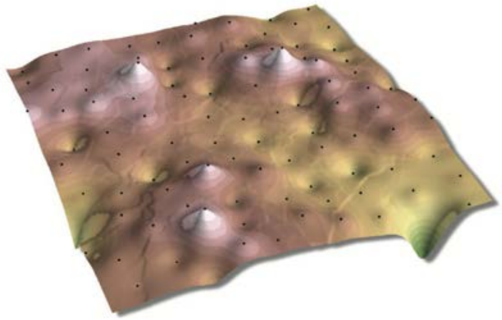
$$w(i) \geq 0, \sum_{i=1}^{N} w(i) = 1$$

TWC 2nd Generation – In Operations Today

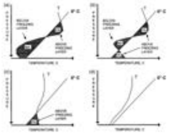
Ensembling concept applied to 4km global grid



Model gradients used to interpolate from model grid resolution

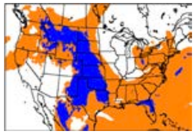


Model data transformed to consumer weather variables



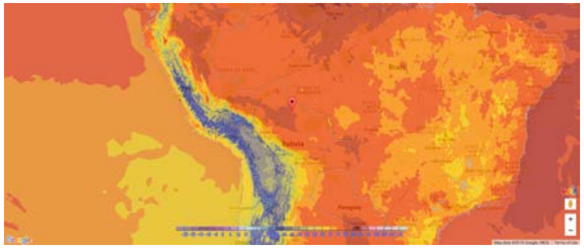
(Bourgouin, 2000) precipitation typing

(Knapp, et al 2006) Thunderstorm Potential Index



(Rasmussen, 1998) visibility and probability of fog

Global observation grid built using a combination of BCDG and Inverse Distance Weighting



Weights and bias adjustments occur via gradient descent (GD), or "nudging"

$$\mathbf{b} \leftarrow \mathbf{b} + \gamma(\mathbf{f} - \mathbf{b} - o)$$

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla \sum_{i=1}^{N} \left(\left(f(i) - b(i) \right) w(i) - o \right)^2$$

Produces 32 million global points, hourly forecasts to 16 days, 17 variables

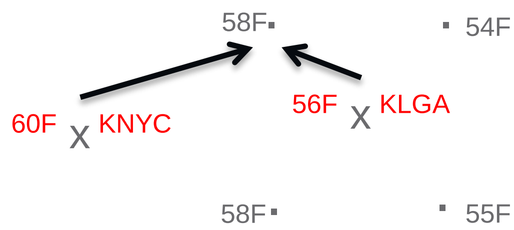
BIG DATA

The Weather Company | The Weather Channel | WU | WSI

TWC 2nd Generation - Limitations

Requires retention of 150 TB of forecast and observation data

Interpolation of observations to grid results in information loss



Interpolated observations and forecasts drive imprecise tuning of weights and biases

Up to 2% decrease in skill!

Inclusion of additional parameters/models requires significant engineering effort

Influence of bad observations is not easily corrected

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

KPDK 132353Z 25004KT 10SM CLR 14/M06 A3018

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

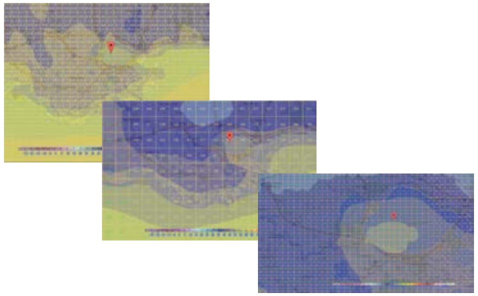
KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

Not straightforward to experiment with new weight and bias adjustment schemes

The Weather Company | The Weather Channel | WSI

TWC 3rd Generation

Preserve native resolution through final forecast creation - don't pre-process models



Weight and bias calculated at observation points using local performance data

KPDK forecast: 4C

KPDK observation:

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

New empirics mathematics – EMA Constrained, Regularized Regression (EMA-CRR)

$$\min_{\mathbf{w}} \mathbf{w}^{T}\Gamma^{agg}\mathbf{w} \quad \text{subj. to } \mathbf{w} \geq 0 \text{ and } \sum w_i = 1$$

Reduced operating costs – lower storage requirements

Influence of bad observations is easily corrected

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

~~KPDK 132353Z 25004KT 10SM CLR 14/M06 A3018~~

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

KPDK 132353Z 25004KT 10SM CLR 04/M06 A3018

Addition of new models is straightforward

+LAMP MOS ✓

+UKMO model ✓

The Weather Company | The Weather Channel | WU | WSI

Constrained Least-Squares Regression

- Compute bias from last M days

$$b(i) = \frac{\sum_{t=1}^{M} \left(f_{-t}(i) - o_{-t} \right)}{M}$$

- Compute error covariance matrix

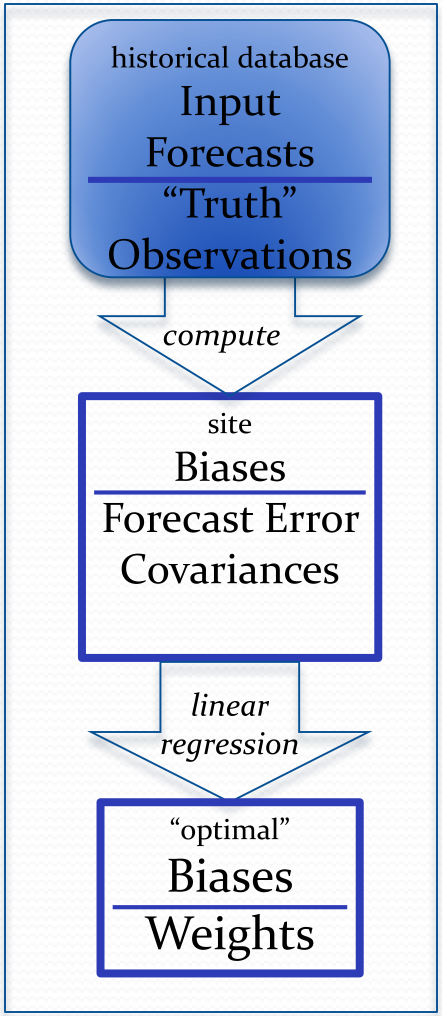
$$\Gamma(i,j) = \sum_{t=1}^{M} \left(f_{-t}(i) - b_{-t}(i) - o_{-t} \right) \left(f_{-t}(j) - b_{-t}(j) - o_{-t} \right)$$

- Choose weights via the quadradic program

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}} \Gamma \; \mathbf{w}$$

$$\text{subj. to } \mathbf{w} \geq 0 \text{ and } \sum w_i = 1$$

(e.g., via MATLAB's 'quadprog')



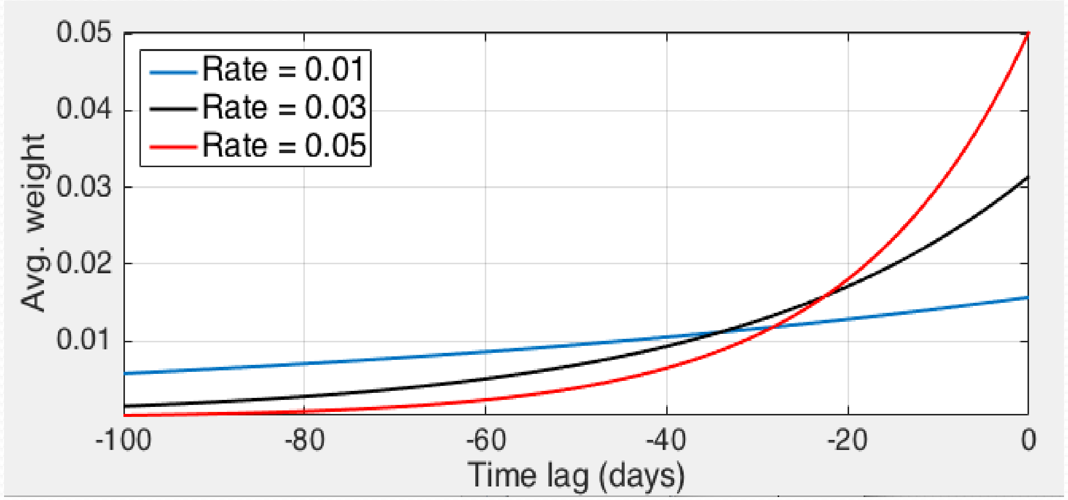
Exponential Moving Average (EMA)

- Compute bias with learning rate *γ* (exponential decay 1 – *γ)*

$$b_0(i) = \frac{\sum_{t=1}^{M} (1-\gamma)^t \left(f_{-t}(i) - o_{-t} \right)}{\sum_{t=1}^{M} (1-\gamma)^t}$$

- Compute Γ with learning rate *η*

$$\Gamma_0(i,j) = \sum_{t=1}^{M} (1-\eta)^t \left(f_{-t}(i) - b_{-t}(i) - o_{-t} \right)\left(f_{-t}(j) - b_{-t}(j) - o_{-t} \right)$$



Regularization

- Construct "aggregate" covariance …

$$\Gamma^{agg} = (1-\alpha)\Gamma_{site} + \alpha\, \Gamma_{neighborhood}$$

- … and add a regularization term ("ridge regression")

$$\Gamma^{agg} = \Gamma^{agg} + \beta\, \mathrm{diag}(\Gamma^{agg})$$

Method Summary: EMA Constrained, Regularized Regression (EMA-CRR)

(1) Dataset: *M* days of forecasts and observations
(2) EMA : biases *b* (rate *γ*) and error covariances Γ (rate *η*)
(3) Γ Aggregation: give neighborhood sites weight *α*
(4) Regularization: inflate Γ diagonal by factor *β*
(5) Constrained solution: solve the quadratic program

$$\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}} \Gamma^{agg} \mathbf{w} \quad \text{subj. to } \mathbf{w} \geq 0 \text{ and } \sum w_i = 1$$

(1) Spreading: interpolate weights and biases from observations to arbitrary forecasts points
(2) Integration: produce consensus forecast via

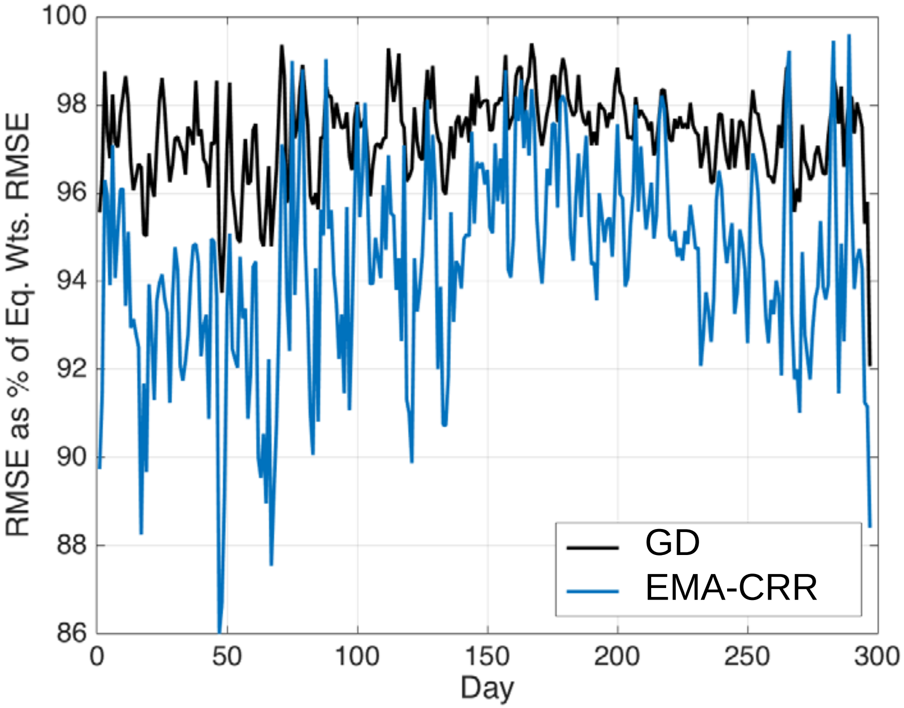
$$F = \sum_{i=1}^{N} \left(f(i) - b(i) \right) w(i)$$

Evaluation

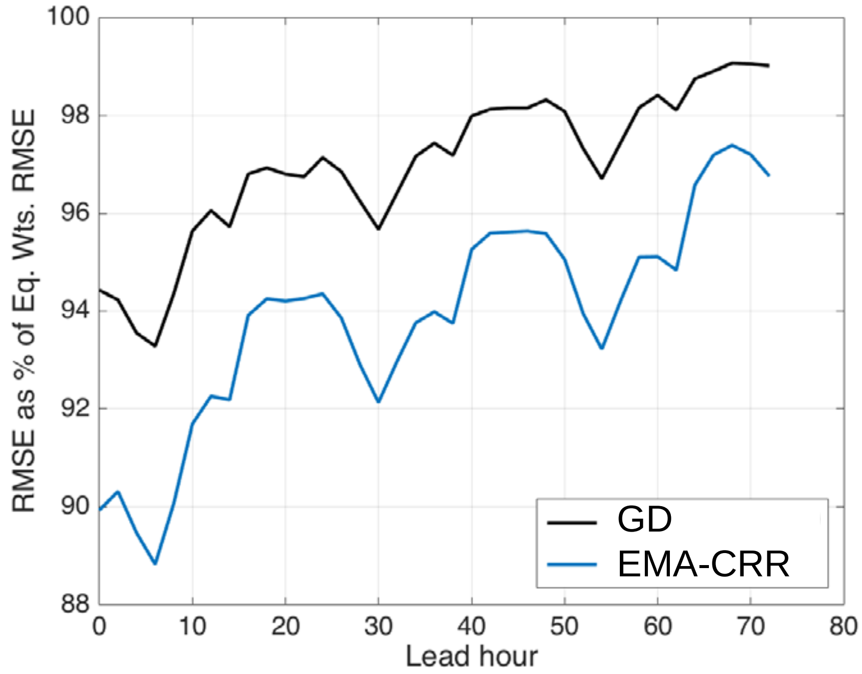
- Temperature at 1200+ CONUS sites
  - 0-72 hour forecasts from 22 NWP model and MOS inputs
  - 0900 UTC forecast generation time
  - November 14, 2014 – January 2, 2016, some missing days
    - "Spin up" November 14 – December 31, 2014
    - Evaluated January 1, 2015 – January 2, 2016
- Parameters chosen by sensitivity tests on selected odd lead hours (not shown)
  - $M = 91$, $\gamma = 0.05$, $\eta = 0.03$, $\alpha = 0.7$, $\beta = 0.1$
- Evaluated RMSE on even forecast lead hours

EMA-CRR RMSE (smaller values better)

Daily RMSE



Lead hour RMSE



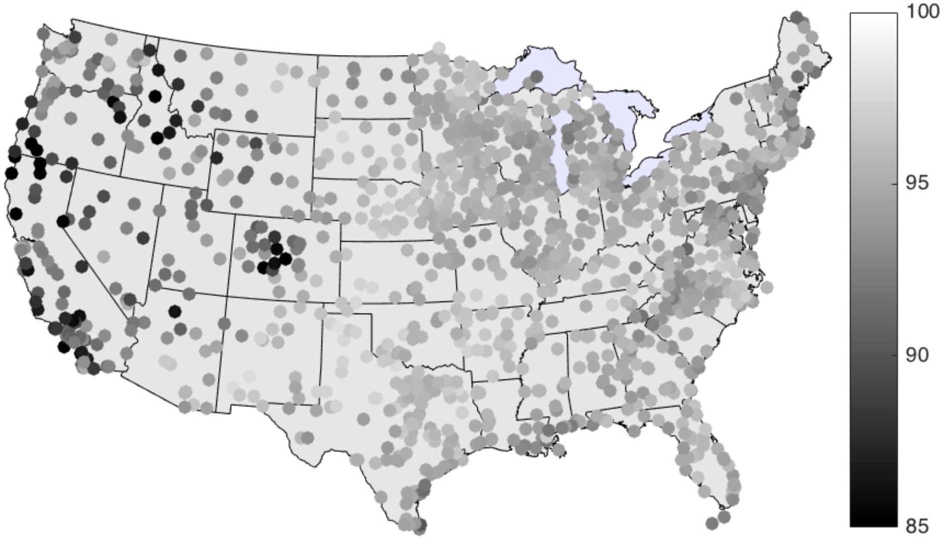
- EMA-CRR and GD consistently better than equal weights
- EMA-CRR better than GD all but 6 days

- EMA-CRR and GD consistently better than equal weights
- EMA-CRR consistently better than GD

EMA-CRR RMSE Site Comparisons

darker shading ➔ greater improvement

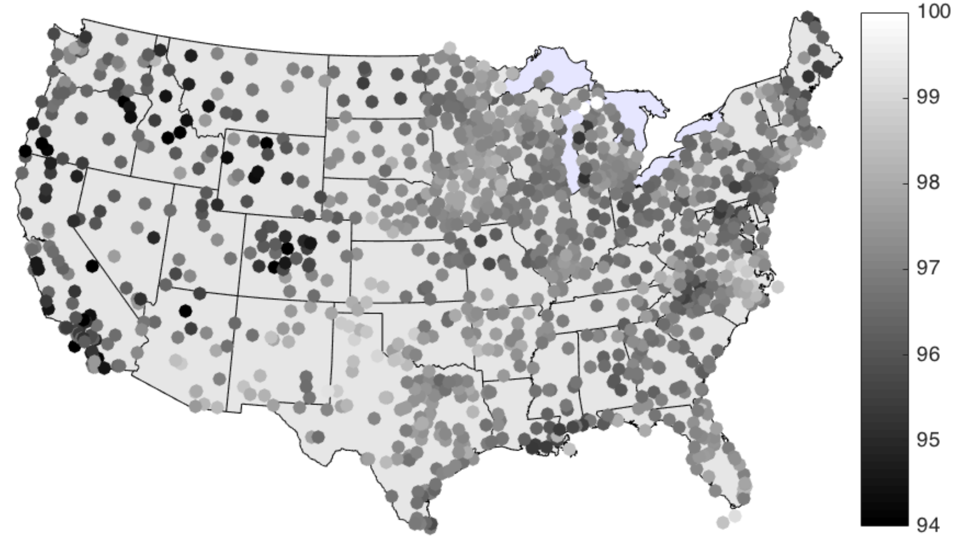
EMA-CRR RMSE as % of Equal Weights RMSE



Median site improvement is 5.1%

(all but 1 site are improved)

EMA-CRR RMSE as % of GD RMSE



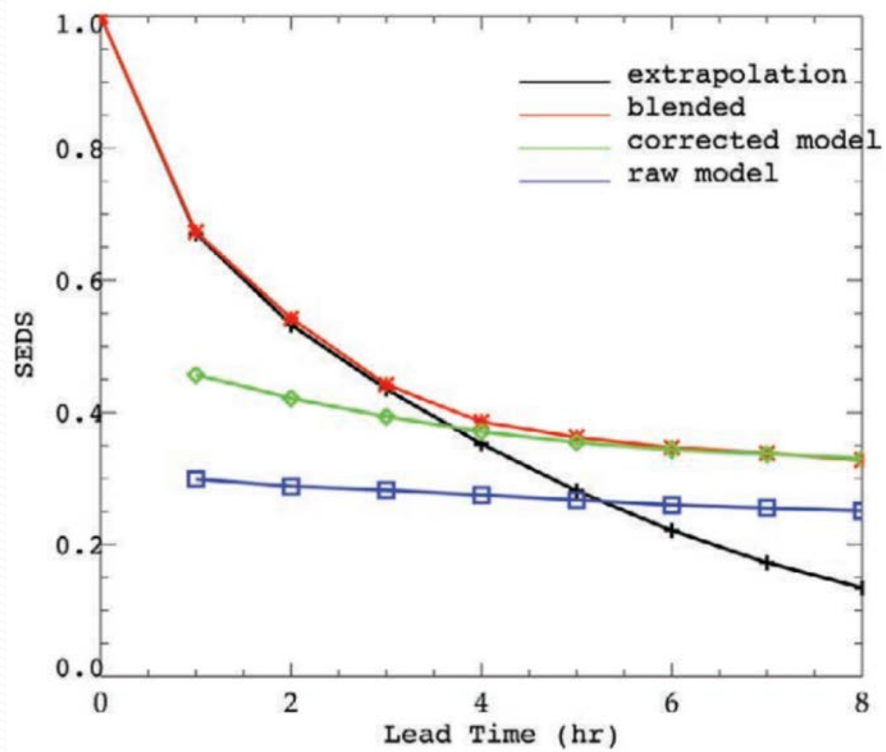
Median site improvement is 3.0%

(all but 4 sites are improved)

Convective Nowcast Model/Radar Blending: Correcting Phase Errors

- FAA "CoSPA" system methodology
- Identify storm "objects" in NWP model and radar VIL
- Compute displacements and their trends
- Adjust model forecast intensity and storm locations
- Compute weighted average with extrapolated radar based on lead-time

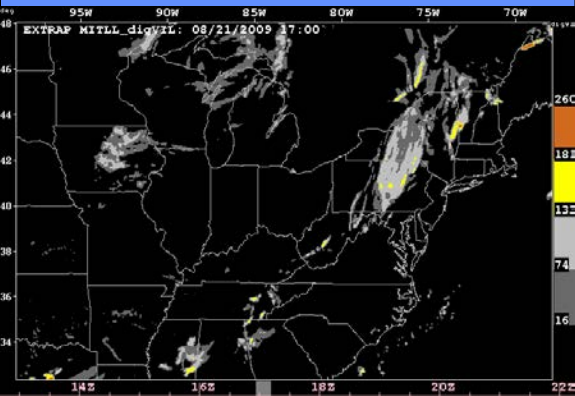
- Could a similar method be applied to temporal phase errors? Or creating "crisper" consensus forecasts?



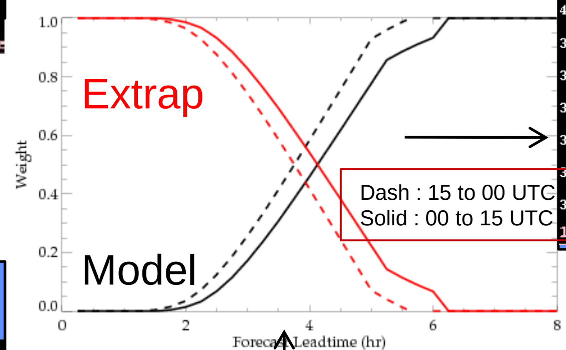
Blending: symmetric extreme dependency score. From Sun et al., *BAMS*, 2014.

Blending Extrapolation and Model

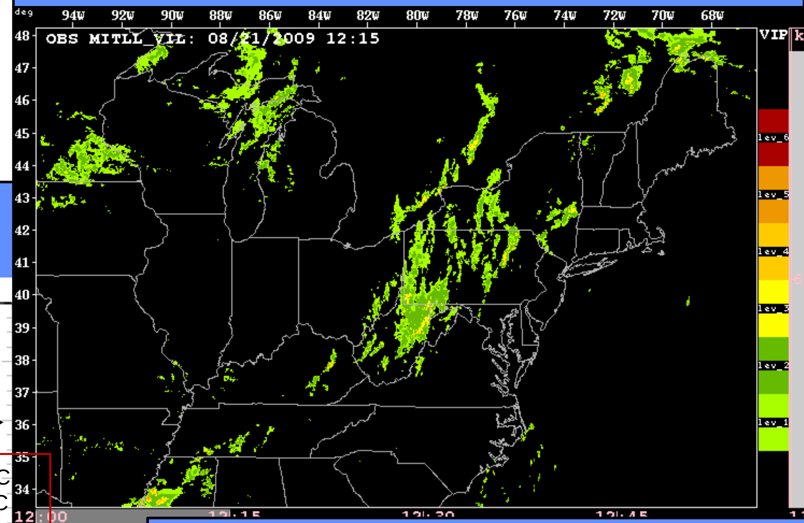
Extrapolation Forecast



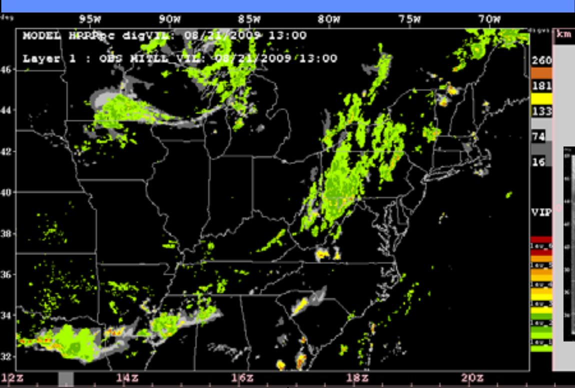
Blending



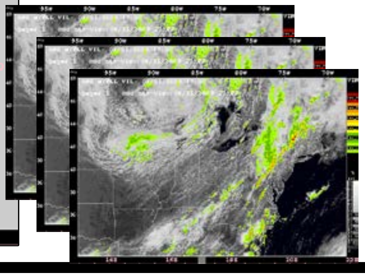
Blended Forecast (12-20 UTC)



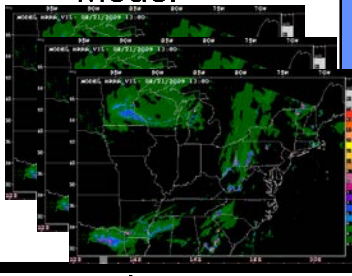
Calibration + Phase Corr



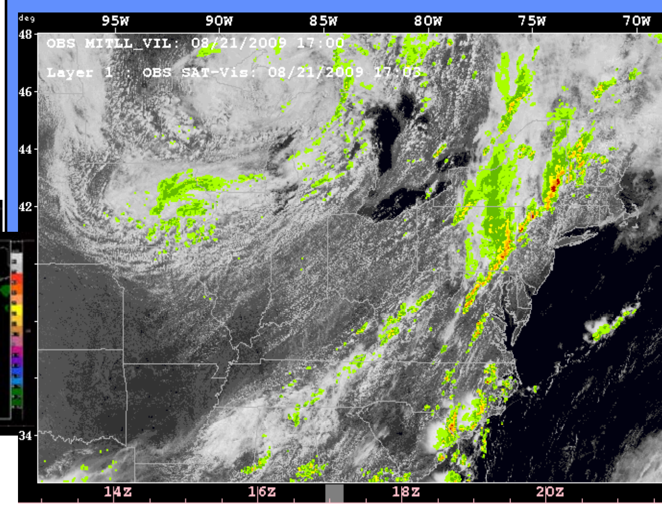
Radar Mosaic



Model



Observations @ 1700 UTC



Slide courtesy of James Pinto, NCAR

Artificial Intelligence Methods “Family Tree”

lots of options to choose from ….

Mathematics

“Soft” computing

Statistics

Optimization Theory

Artificial Intelligence / Computational Intelligence

Machine Learning / Stat. Learning / Data Mining

Expert Systems

Supervised Learning

Unsupervised learning

Fuzzy Logic

Decision Trees

PR
Poly. Regression

SVM
Support Vect. Machine

DRT
Decision/Reg. Trees

GA
Genetic Algorithm

SA
Simulated Annealing

RL
Reinforc. Learning

NBC
Naïve Bayes Class

ANN
Art. Neural Networks

KNN
k-Nearest Neighbors

RF
Random Forests

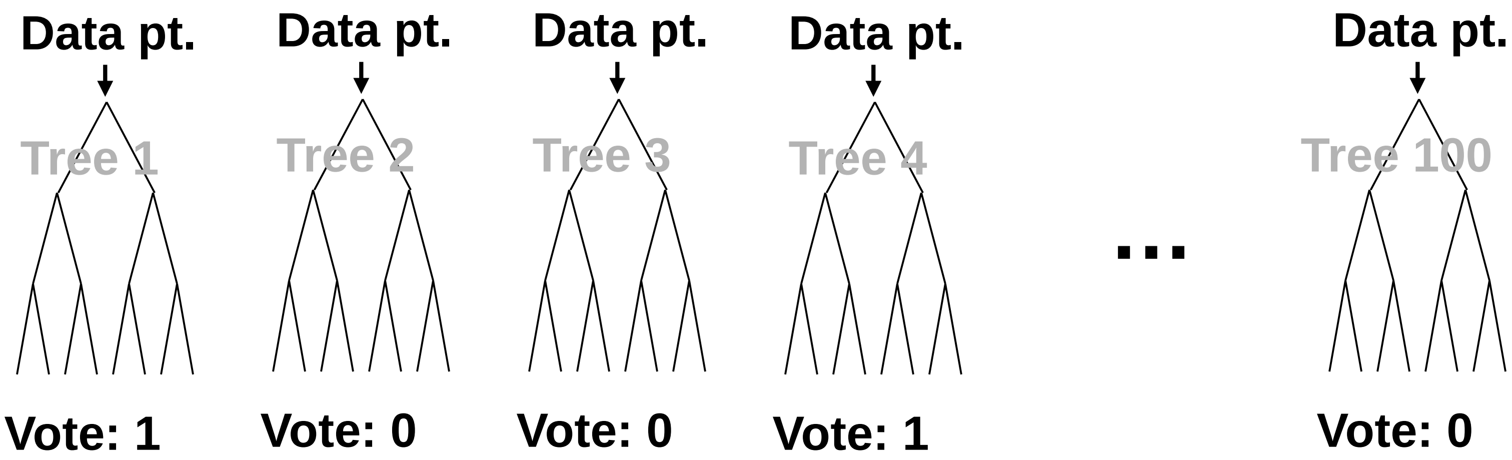
PCA
Principal Comp. Anal.

SOM
Self-organ. map

FCM
Fuzzy *c*-means clust.

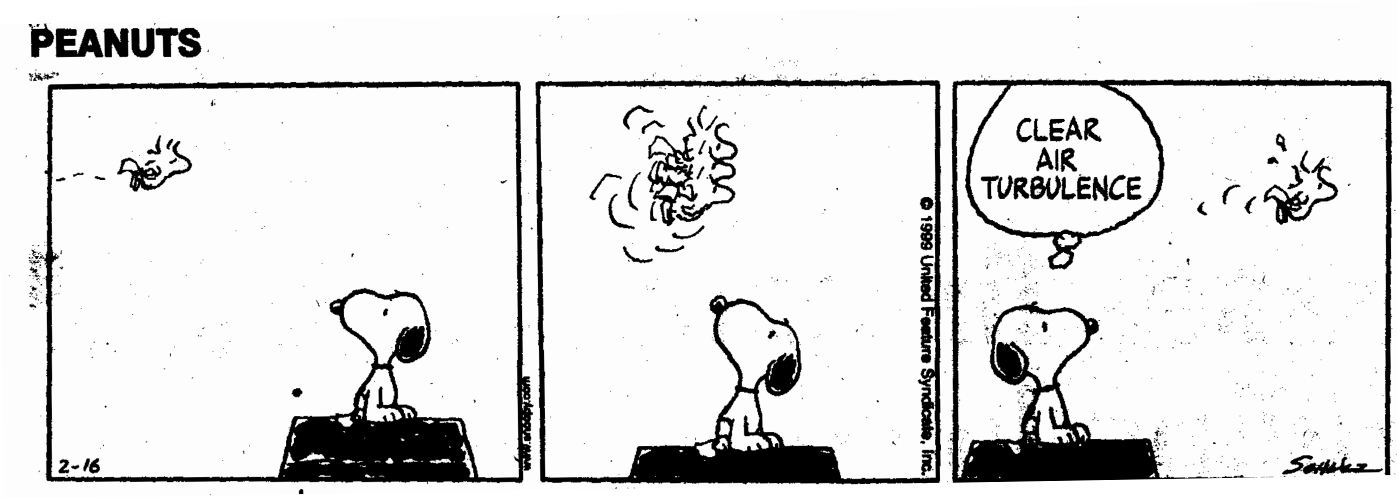
Decision Tree Method: Random Forest (RF)

- Very general method for creating predictions
- Uses "training set" of predictor variables and associated labels (for classification) or values (for regression)
  - Tree strained with random subsets of data and predictors
- Produce estimates of predictor "importance"
- "Votes" are calibrated to create reliable probabilities or deterministic predictions



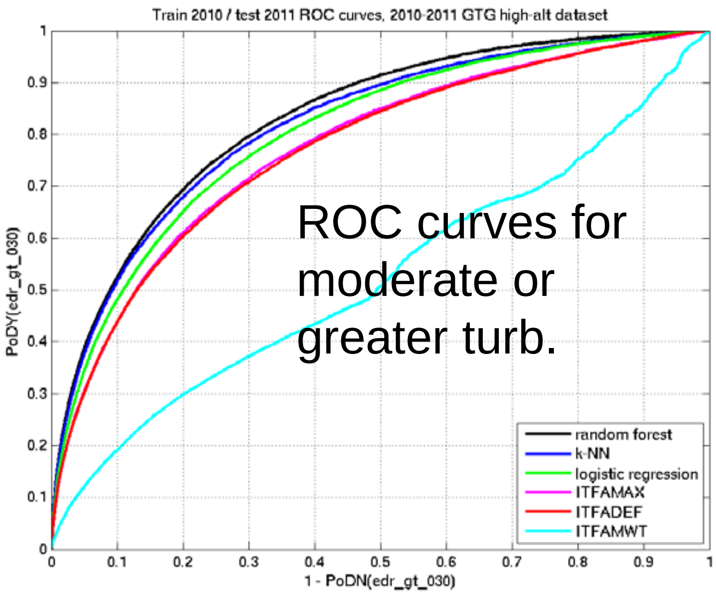
E.g., 40 votes for "0", 60 votes for "1"

En-route Aviation Turbulence



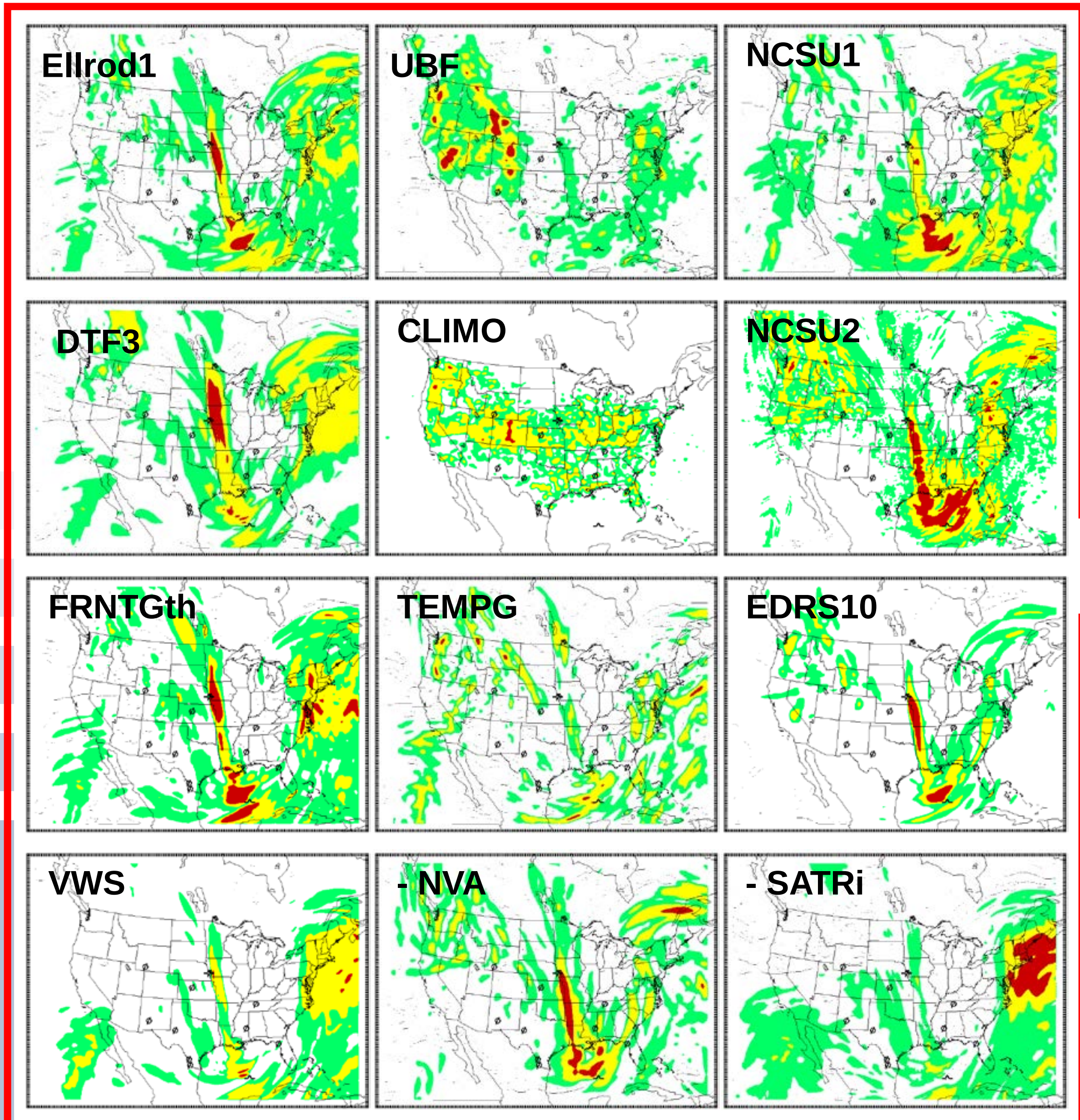
En-route Aviation Turbulence

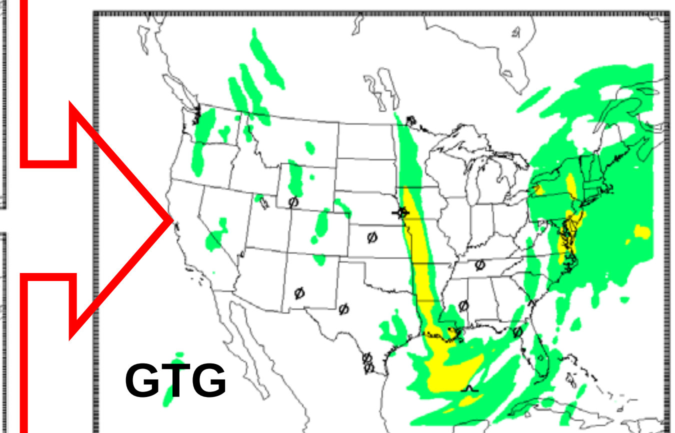
- NWP models don't resolve aircraft-scale turbulence (~10 to 100+ m)
  - TKE shows poor correlation with aircraft observations
- Compute a set of turbulence "diagnostics" from 3-D model output
  - Richardson number, vertical wind shear, structure function EDR, …
  - + Proximity to VIL, echo top, cloud top height contours, …
- Train using aircraft turbulence observations as "truth"
  - Under-sample null turbulence cases, then recalibrate trained model
  - Train separate combinations for different altitude bands



	AUC	Max CSI	Max TSS
Random Forest	0.839 (0.830-0.849)	0.174 (0.169-0.179)	0.520 (0.499-0.541)
K-Nearest Neighbors	0.825 (0.818-0.832)	0.165 (0.163-0.167)	0.499 (0.482-0.514)
Logistic Regression	0.813 (0.805-0.822)	0.156 (0.149-0.162)	0.477 (0.461-0.496)
GTG_max	0.783 (0.775-0.791)	0.135 (0.132-0.137)	0.430 (0.418-0.443)
GTG_cat	0.780 (0.772-0.789)	0.134 (0.131-0.136)	0.427 (0.412-0.441)

Merging a variety of turbulence diagnostics into “Graphical Turbulence Guidance”

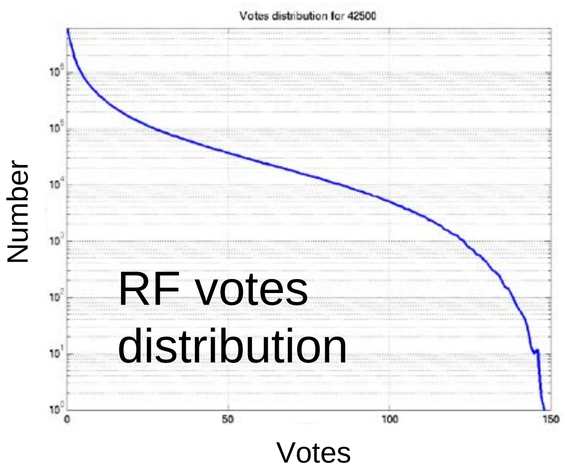


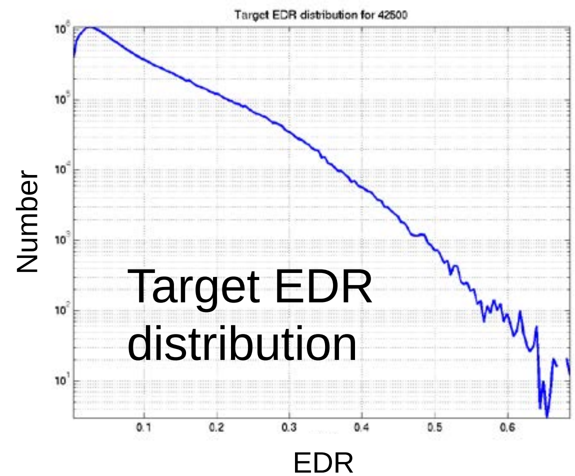


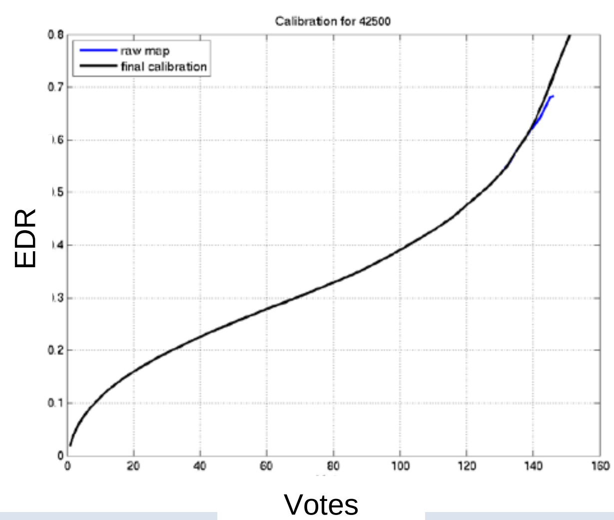
6 h forecast valid at 5 Feb 2006 00Z

RF Calibration to EDR

- RF models trained on even Julian days and evaluated on odd, and vice-versa, 64 cross-evaluations in all
- Calibration maps from vote distribution to a target EDR distribution





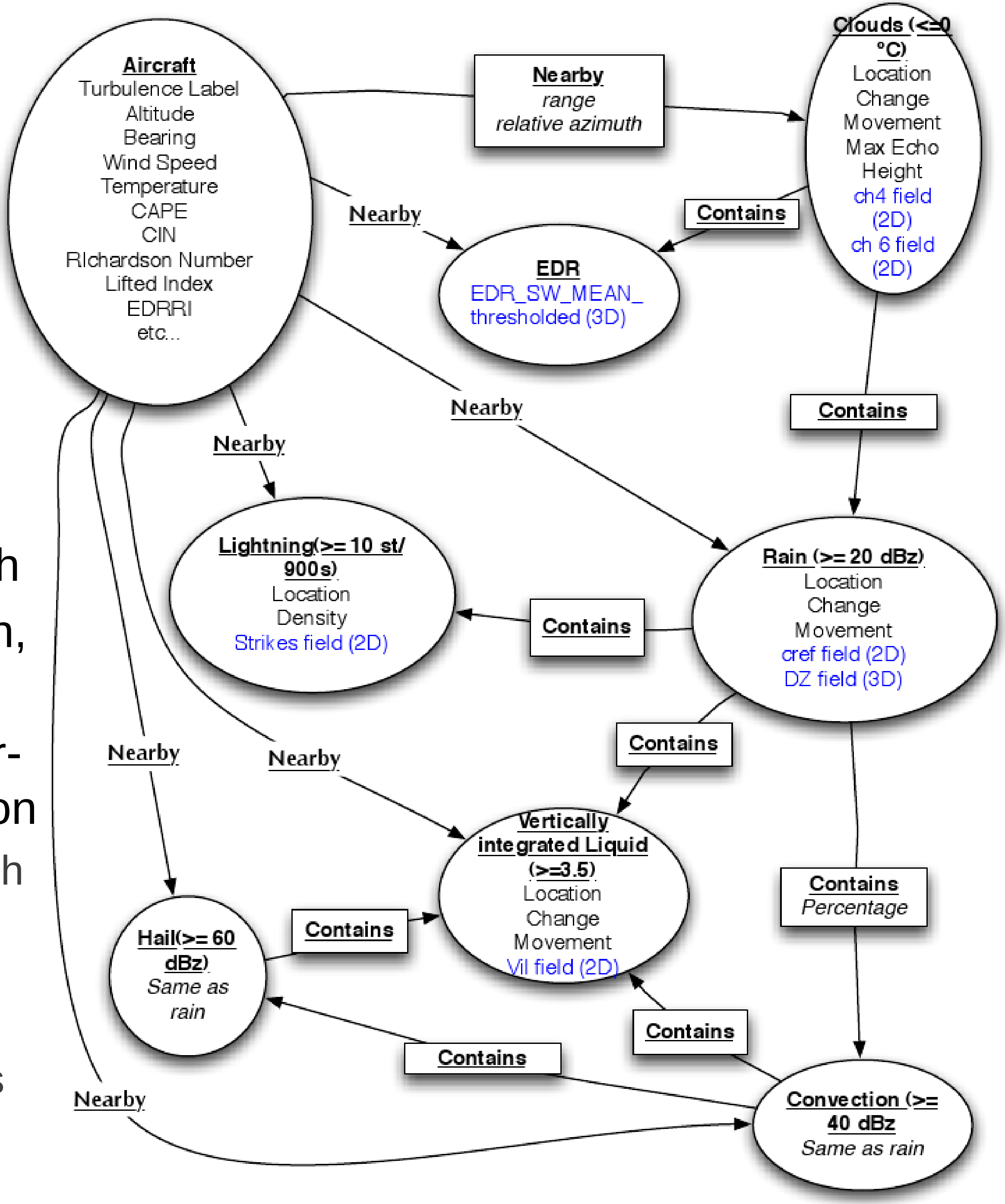


Calibration
for 35 - 50 kft

See Williams, *Machine Learning*, 2014

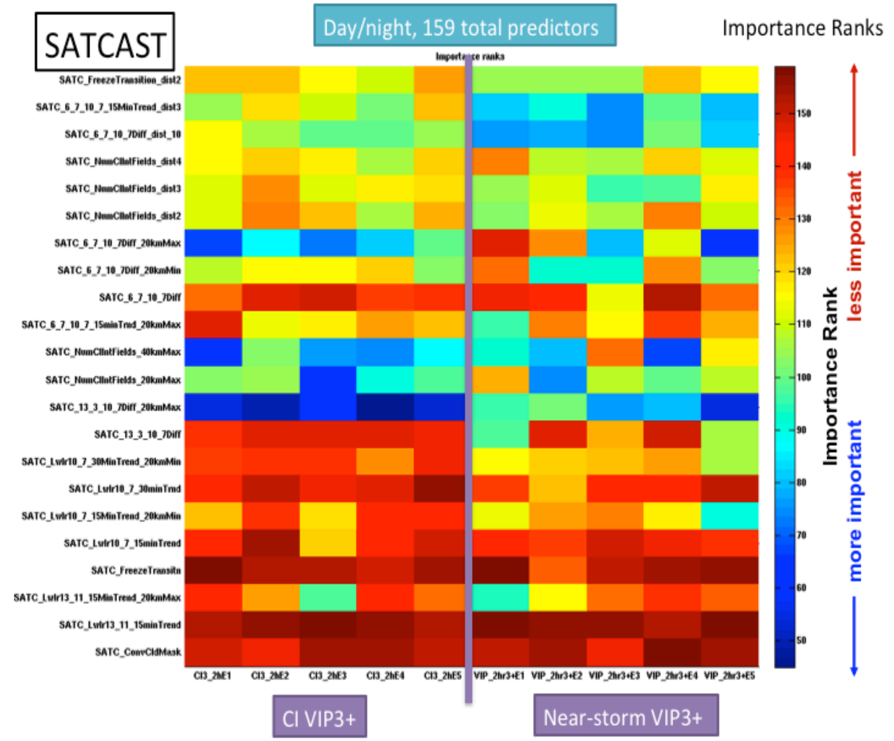
Spatio-temporal relational probability trees (SRPTs) and Random Forests (SRRFs)

- Object relational approach
- Work with Amy McGovern, DJ Gagne at OU
- Example schema for near-storm turbulence prediction
  - Method searches through methods and thresholds for useful “splits” when building trees
  - trees output probabilities



Forecasting Convection

- Used random forest methodology to predict VIL at each pixel
- Truth: radar VIL advected backwards to be coincident with antecedent observations
- Predictors: model and satellite data
- Used importances to choose regimes, predictors
- Outperforms competitive nowcasts

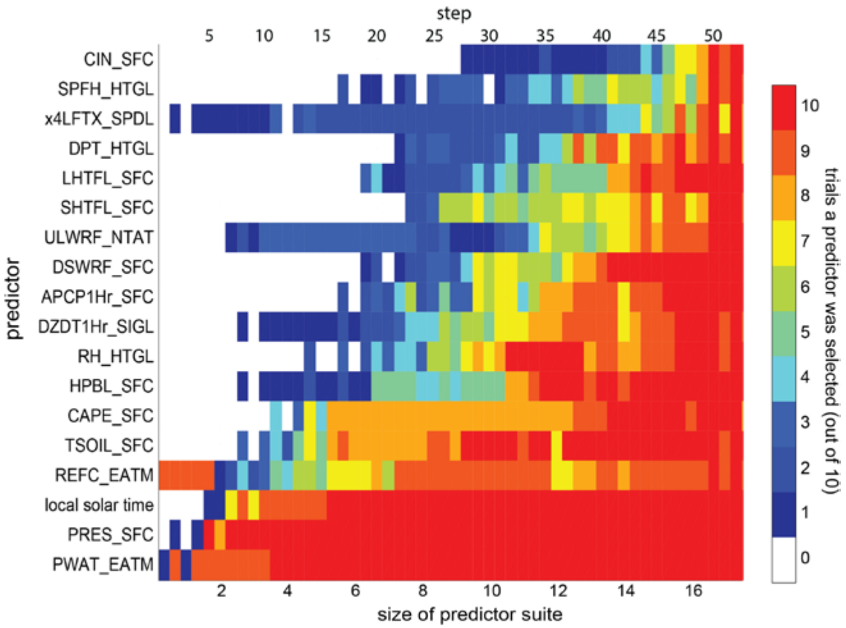


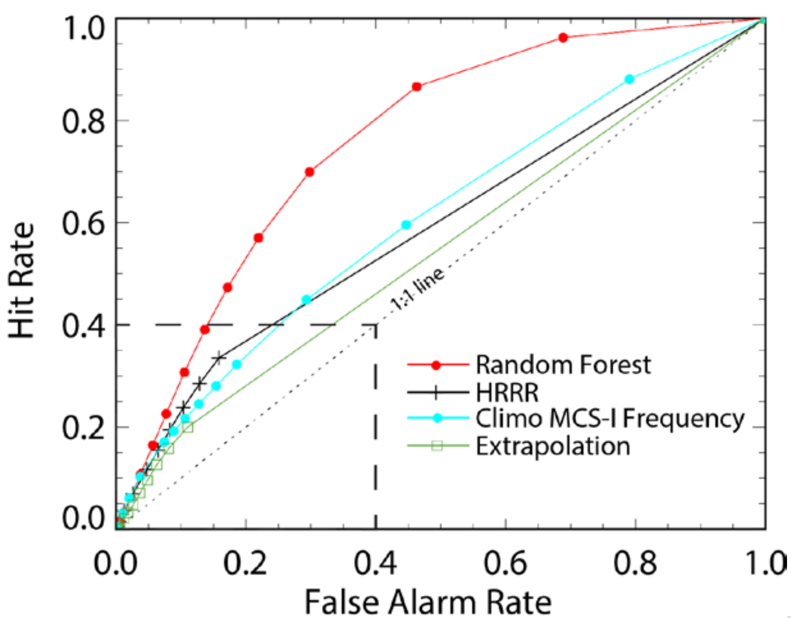
Predicting Convective Initiation	Max CSI	Max TSS	AUC
2h simple extrapolation	0.005 ± 0.002	0.17 ± 0.05	0.60 ± 0.03
CoSPA (2h)	0.012 ± 0.005	0.12 ± 0.03	0.56 ± 0.02
LAMP 1-3h (2hr)	0.023 ± 0.006	0.56 ± 0.03	0.83 ± 0.01
2h RF	**0.032** ± 0.011	**0.68** ± 0.02	**0.91** ± 0.01

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

Forecasting MCS Initiation

- MCSs are hazardous to the public and significant for aviation
- MCS-I: 125-km buffer around VIL > 3.5 kg m^{-2} with extent ≥ 100 km (allowing gaps ≤ 10 km)
- Dataset undersampled no-MCS-I cases
- RF used to create 2-hour MCS-I forecasts based on (smoothed) HRRR, satellite, and extrapolated radar
- Forward-backward variable selection used to select predictors





From Ahijevych et al., *WAF*, 2016

M5’ Trees (Quinlan 1992): Cubist

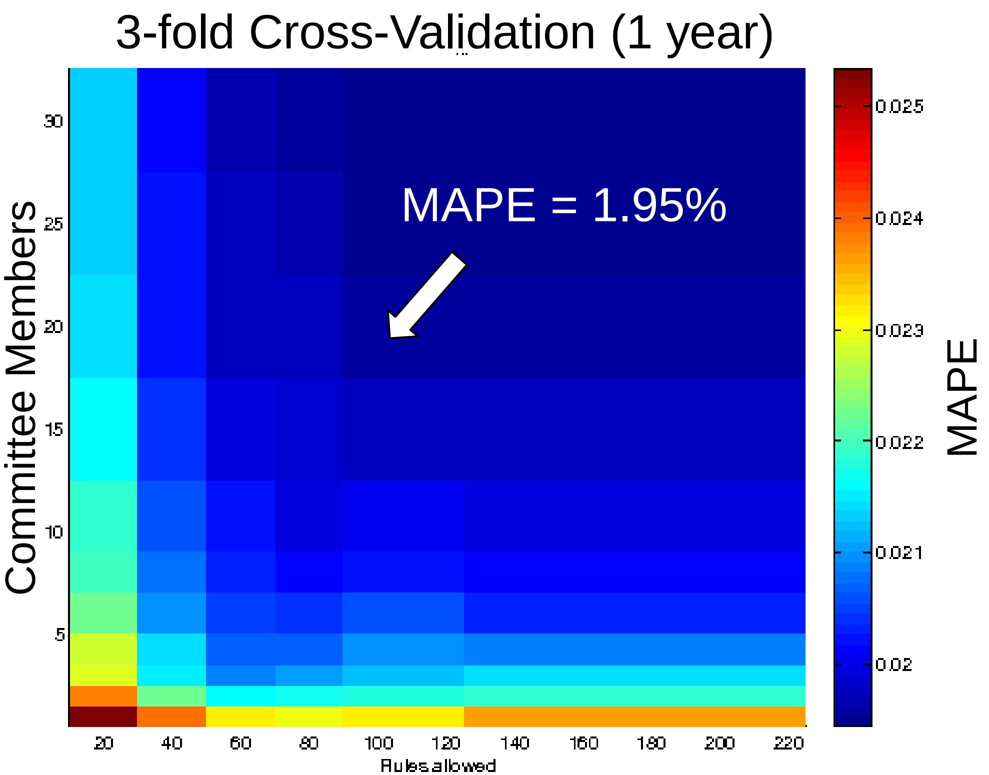
- Regression using categorical and continuous predictors
- Automatically creates “rules”, with multi-linear predictive function for each
  - i.e., can identify significant “regimes” for prediction
- E.g. for computing electrical load, simplified:

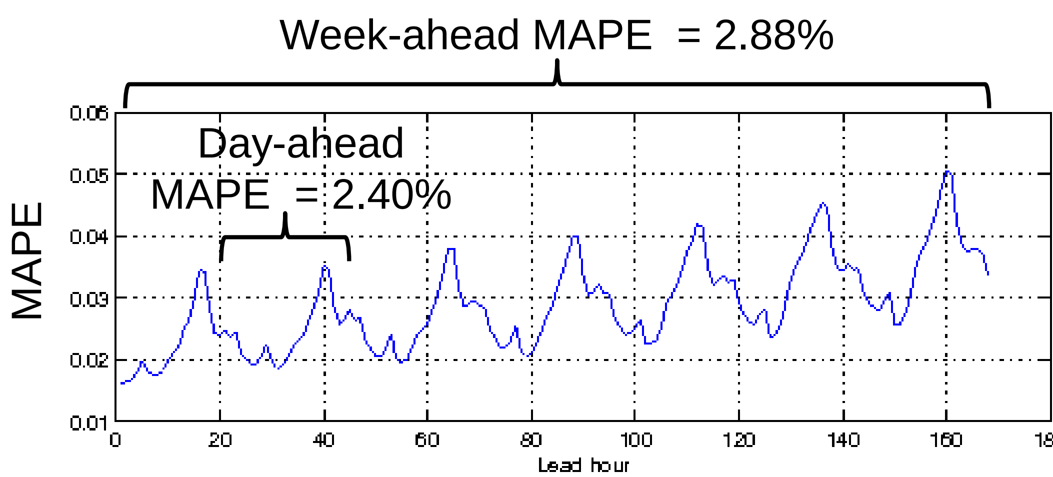
```
if     LocalHour in {13, 14, 15, 16}
       UtcDayOfWeek in {0, 1, 2, 3, 4}
       Temperature >= 18.502
then   obLoad = 3.051 + 1.029 Temperature
                  - 0.0018 Cloud_cover
```

- Aggregates applicable rules to make predictions
- Can be run in “committee” mode

Cubist for Net Load Forecasting

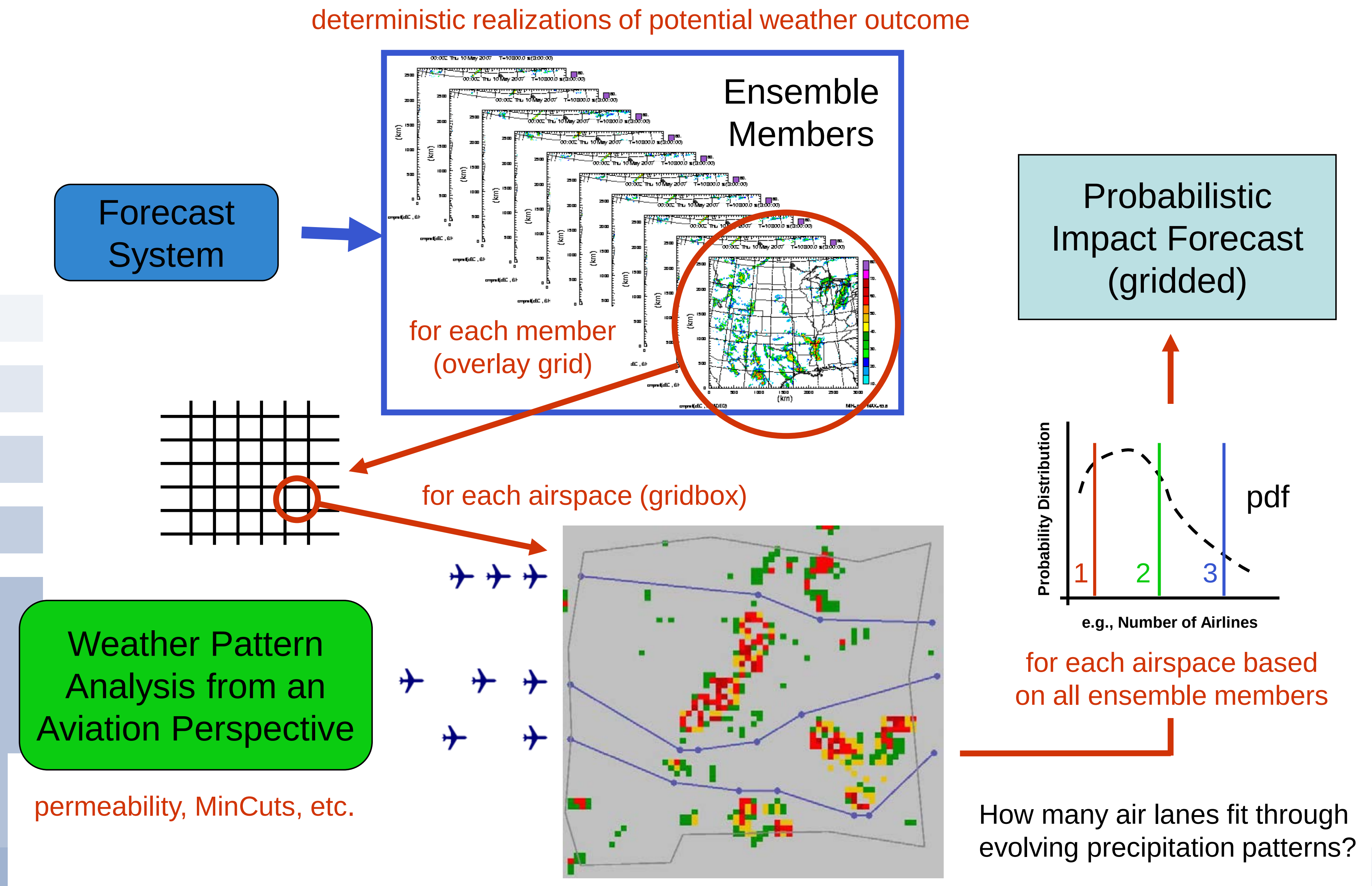
- Predictors include Month, Day of Week, Holiday, Local Hour, Season, temperature, probability of precip, dew point, cloud cover, EMA temperatures …
- Trained daily on previous load obs, applied to following week's forecast
- Load obs dataset begins 1 Sep 2012
- Results from 1 Oct 2013 – 30 Sep 2014





Post-processing Ensembles for Decision Support

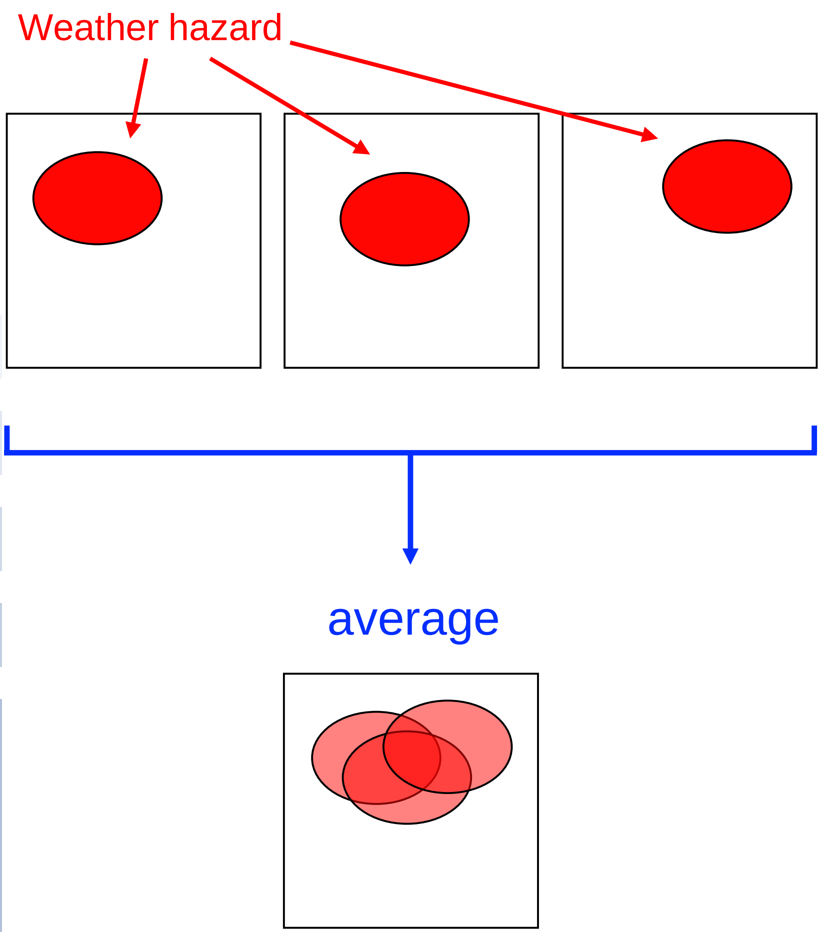
slide courtesy of Matthias Steiner, NCAR



Post-processing Ensembles for Decision Support (2)

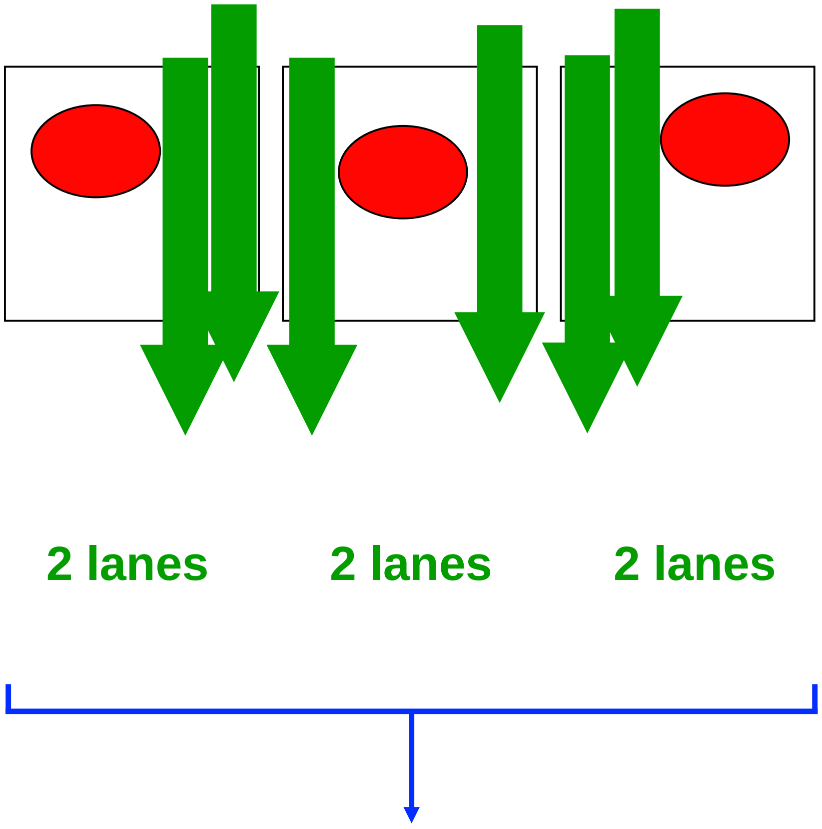
slide courtesy of Matthias Steiner, NCAR

Old way . . .



How many air lanes?

New way . . .



Most likely 2 air lanes . . .

Crowdsourcing

- Contests for methods development
  - Dataset preparation is a huge barrier to methods development
  - Methods comparisons requires using exactly the same data
  - Contest model:
    - Publish datasets and frame learning goal(s), verification metrics
    - Evaluate methods based on holdout datasets
    - Declare a winner: cash prize will encourage entries
    - E.g., AMS AI contests, Kaggle
- “Markets” for real-time forecasting
  - Prediction markets have proven valuable in many domains
  - Fantasy sports, why not fantasy weather!
    - Amateur forecasters put their money on different outcomes
    - Consensus is likely a good synthesis of available information

Summary and Conclusion

- The Weather Company is implementing a new-generation forecast
  - Point-based enables longer obs and model history
  - New EMA-CRR method produces more skillful weights
- Appropriate post-processing may be use-specific
  - Phase error correction
  - AI methods to accommodate flexible predictors and predictands
  - Calibration to event probabilities
  - Decision support applications may ensemble *outcomes* from weather response model run on individual realizations or exemplars
- Crowdsourcing for model development or even forecasting
- There is unlikely to be a single "best method"
  - User requirements and metrics differ
  - Ensemble of methods may be useful