Implementation and Evaluation of a Mesoscale Short-Range Ensemble Forecasting System Over the Pacific Northwest

Eric P. Grimit

Advisor: Dr. Cliff Mass

Department of Atmospheric Sciences, University of Washington

Seattle, Washington

Outline

- **Ensemble Forecasting Background**
  - **Theory**
  - **Potential Benefits**
  - **Applications**
- Motivation for this Research
- The Approach
- Results
- Conclusions
- Future Work

Paradigm Shift

Single atmospheric prediction

Deterministic view of numerical weather prediction (NWP):

- Obtain BEST POSSIBLE initial state
- Use the BEST POSSIBLE NWP model
- Run model at the HIGHEST GRID RESOLUTION POSSIBLE
- Forecasters rely on this ALL or NOTHING forecast

↓

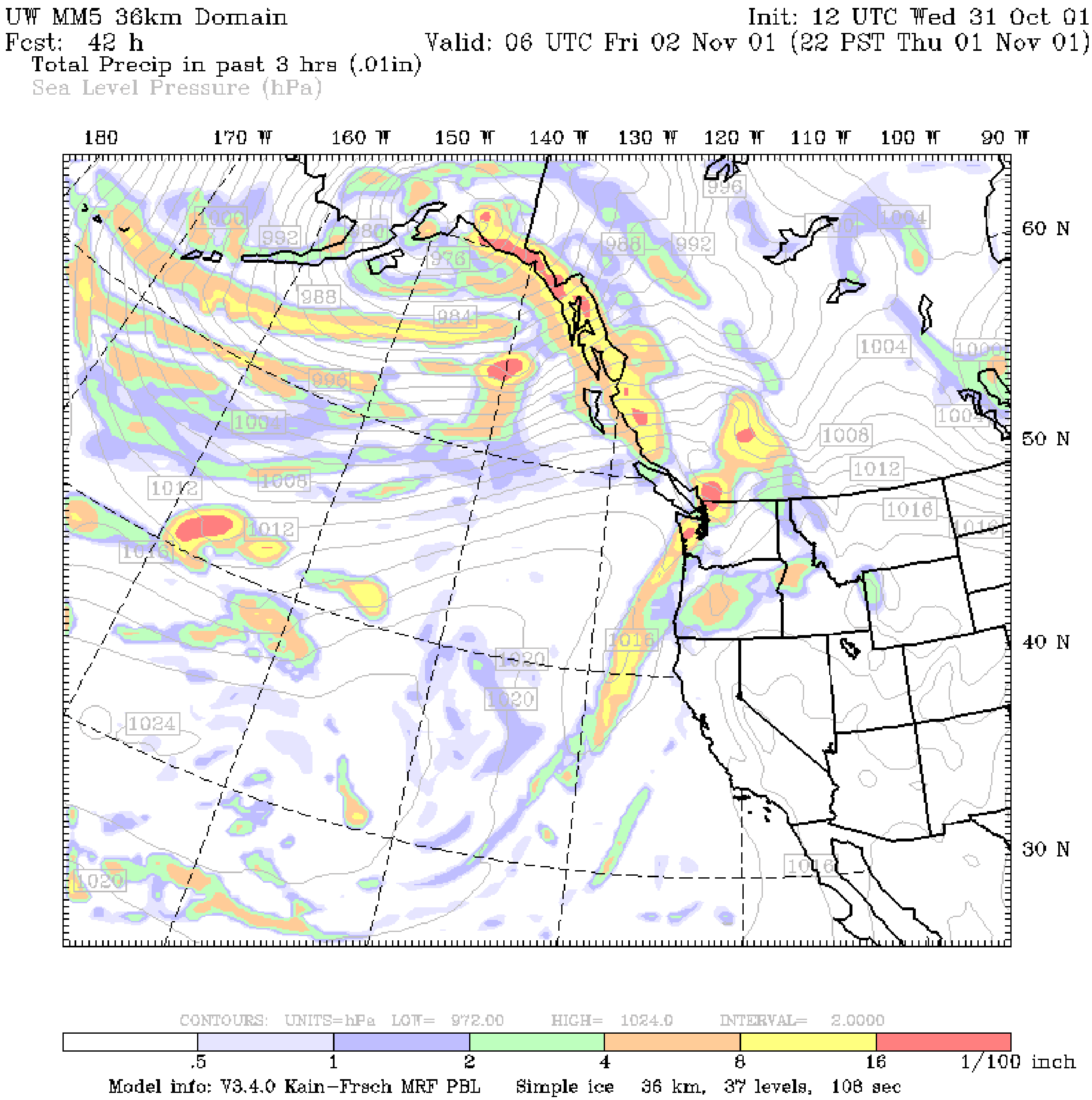
Multiple atmospheric predictions

Ensemble Forecasting

Probabilistic view of NWP:

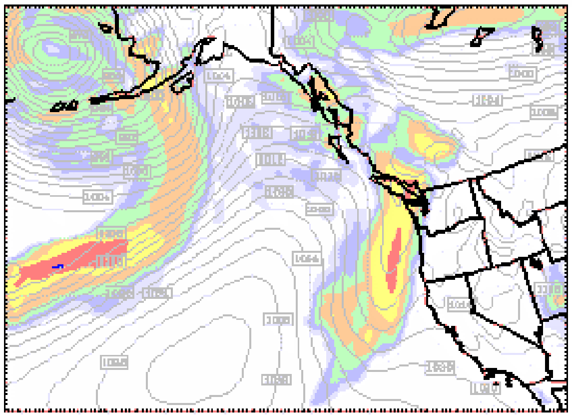
- Obtain a RANGE of POSSIBLE initial states
- Use a NUMBER of VALID models or model configurations
- Run model at lesser grid resolution
- Forecasters get a RANGE of POSSIBLE forecasts → probability forecasts

Deterministic MM5 forecast example:

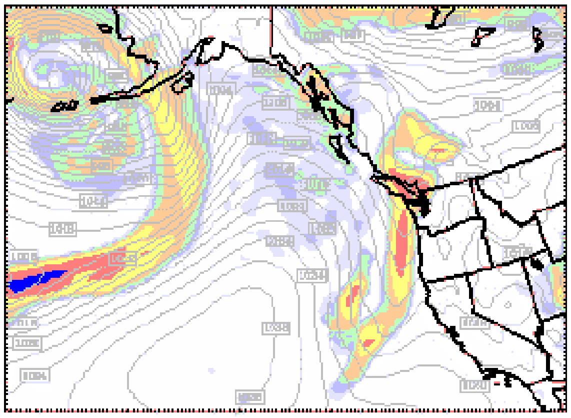


What if we had MULTIPLE MM5 forecasts available to us?

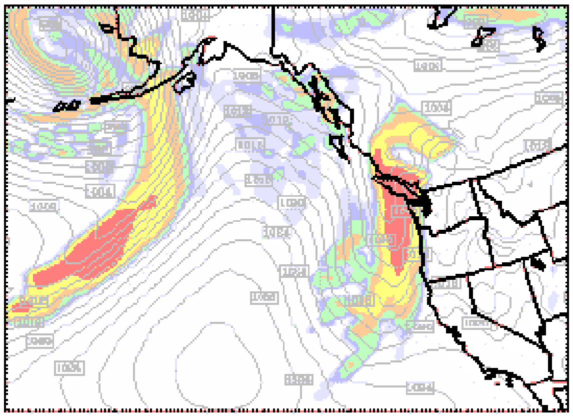
TOT MEAN-MM5 ENSM 49-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01



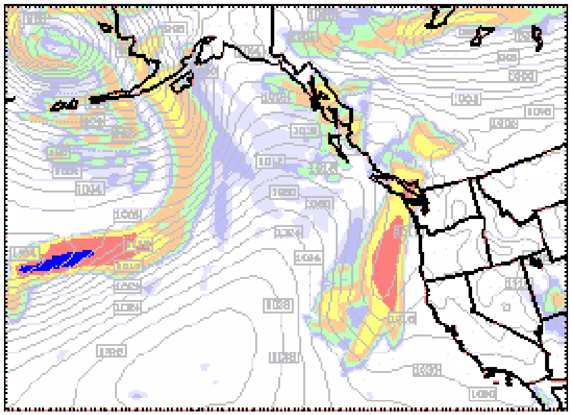
AVN-MM5 ENSM 36km 48-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01



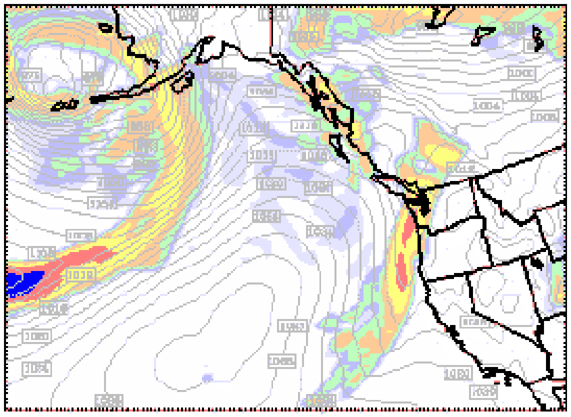
CMCGEM-MM5 ENSM 36 48-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01



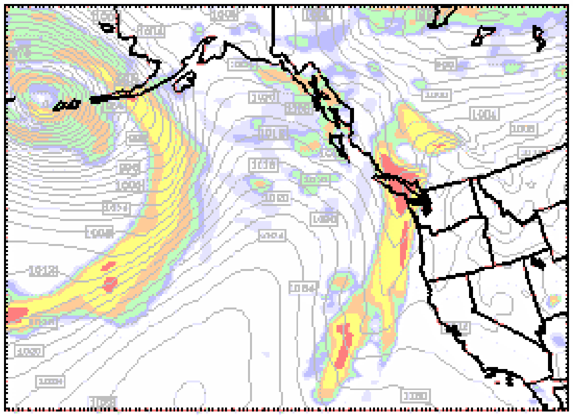
ETA-MM5 ENSM 36km 48-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01



MRF-MM5 ENSM 36km 48-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01



NOGAPS-MM5 ENSM 36 48-hr Fcst Valid: 00 UTC MON 29 JAN 01
Initialized: 00 UTC SAT 27 JAN 01 16 PST SUN 28 JAN 01

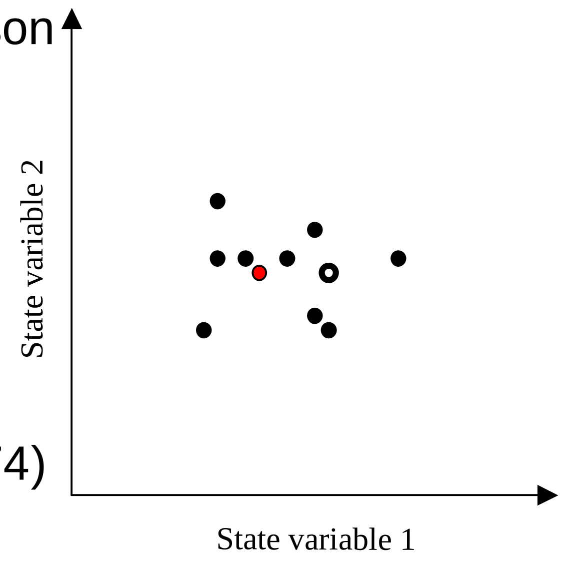


Ensemble Theory

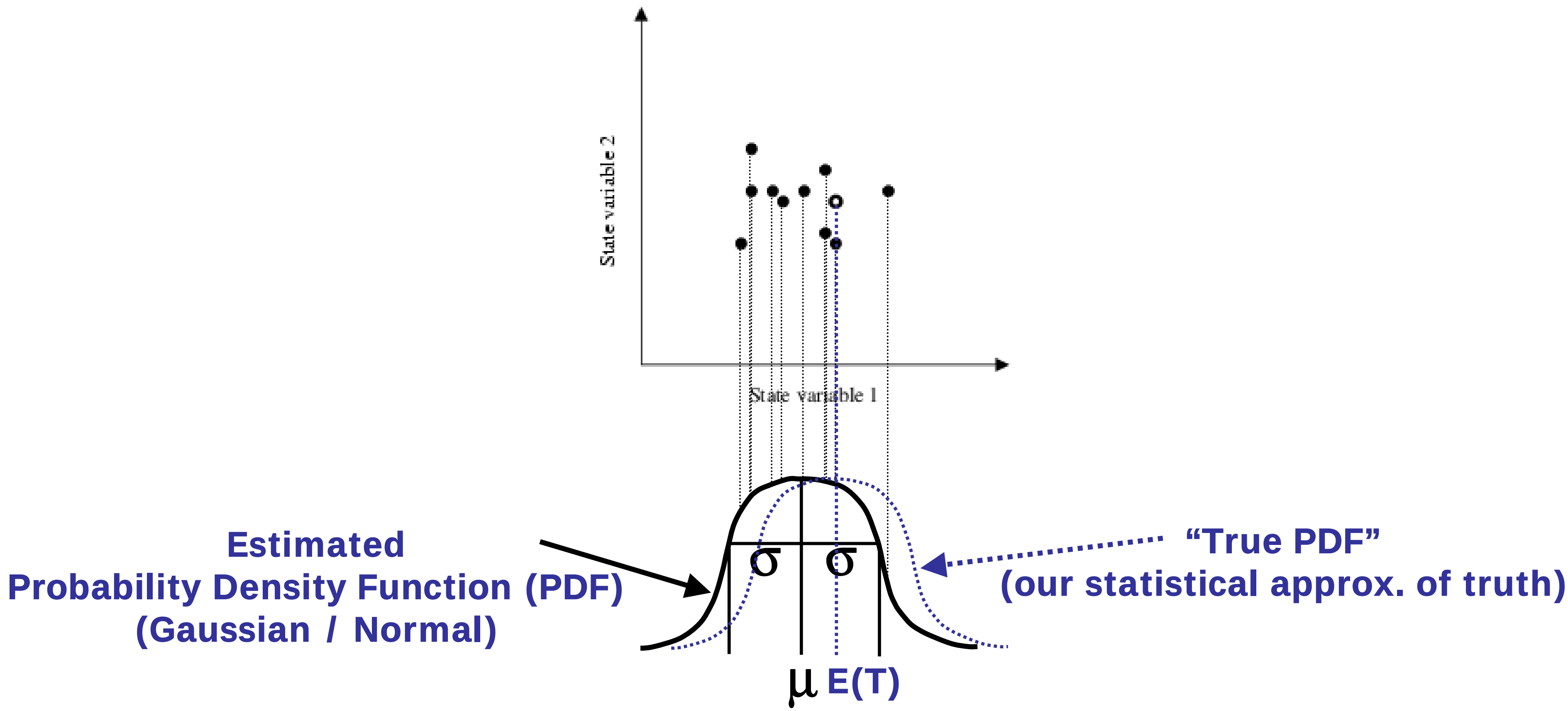
- Atmospheric predictability is limited by:
  - **Analysis error [initial condition (IC) uncertainty] – (Lorenz 1963, 1969)**
  - **Model error [model physics uncertainty]**

- Ensemble forecasting (EF) is an approximation to stochastic-dynamic prediction (Epstein 1969; Gleeson 1970)
  - Prognostic equations are modified to account for IC errors
  - Uncertainty carried throughout the model forecast
  - Yields a probability distribution of solutions
  - Not yet practical for implementation on computers

- Instead, select a finite number of specific ICs (Leith 1974)
  - Monte Carlo (random perturbations from a control)

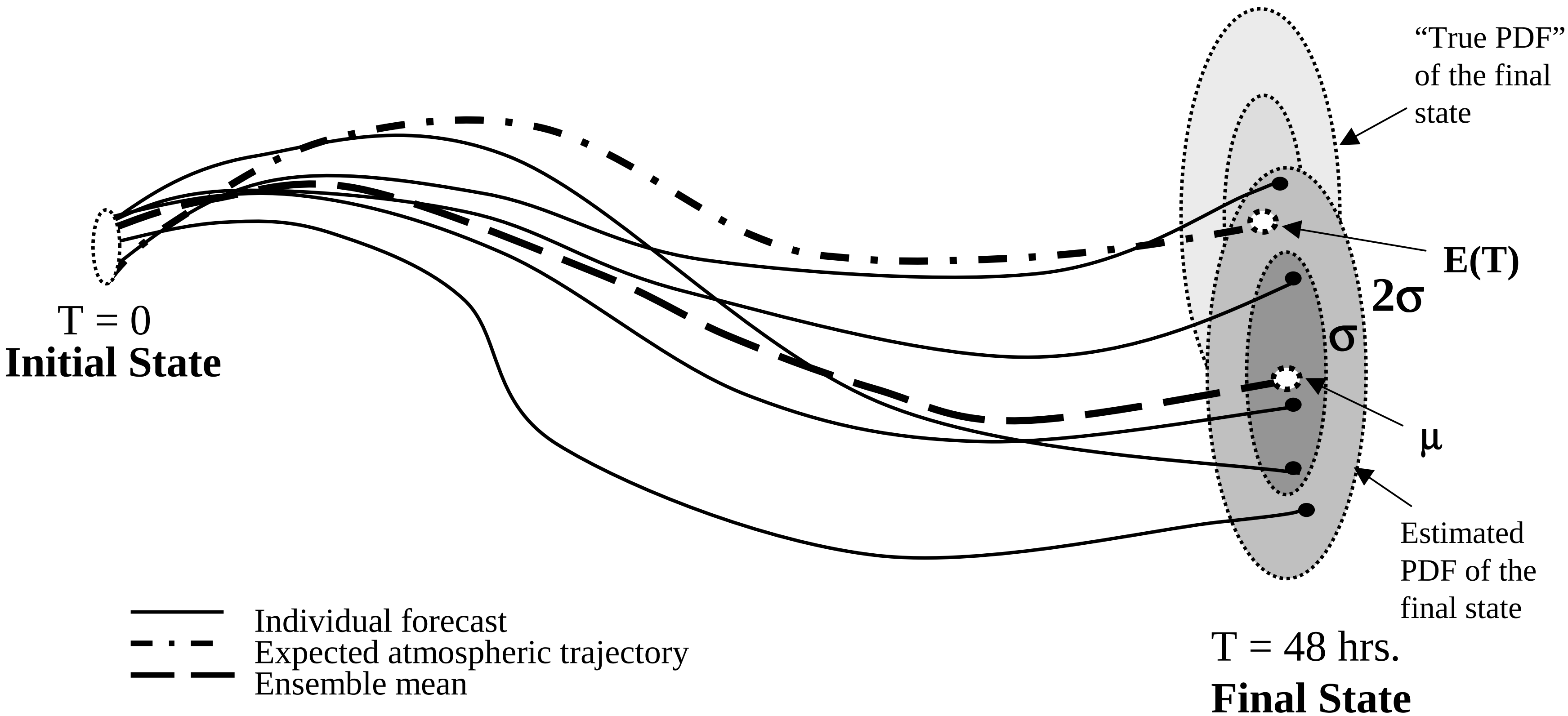


How many Monte Carlo ICs are needed?



Assuming that the ICs are independently selected from the same distribution as truth...

- Very large number required to converge toward the true PDF
- Tradeoff dilemma between computer power and running a large enough ensemble



Ensemble-mean Skill

$$\lim_{t \to \infty} \bar{V} = \frac{1}{2}(1+M^{-1})\tilde{V}$$

$\bar{V}$ = error variance of ensemble mean (μ)

M = number of ensemble members (5)

$\tilde{V}$ = error variance of each ensemble member forecast

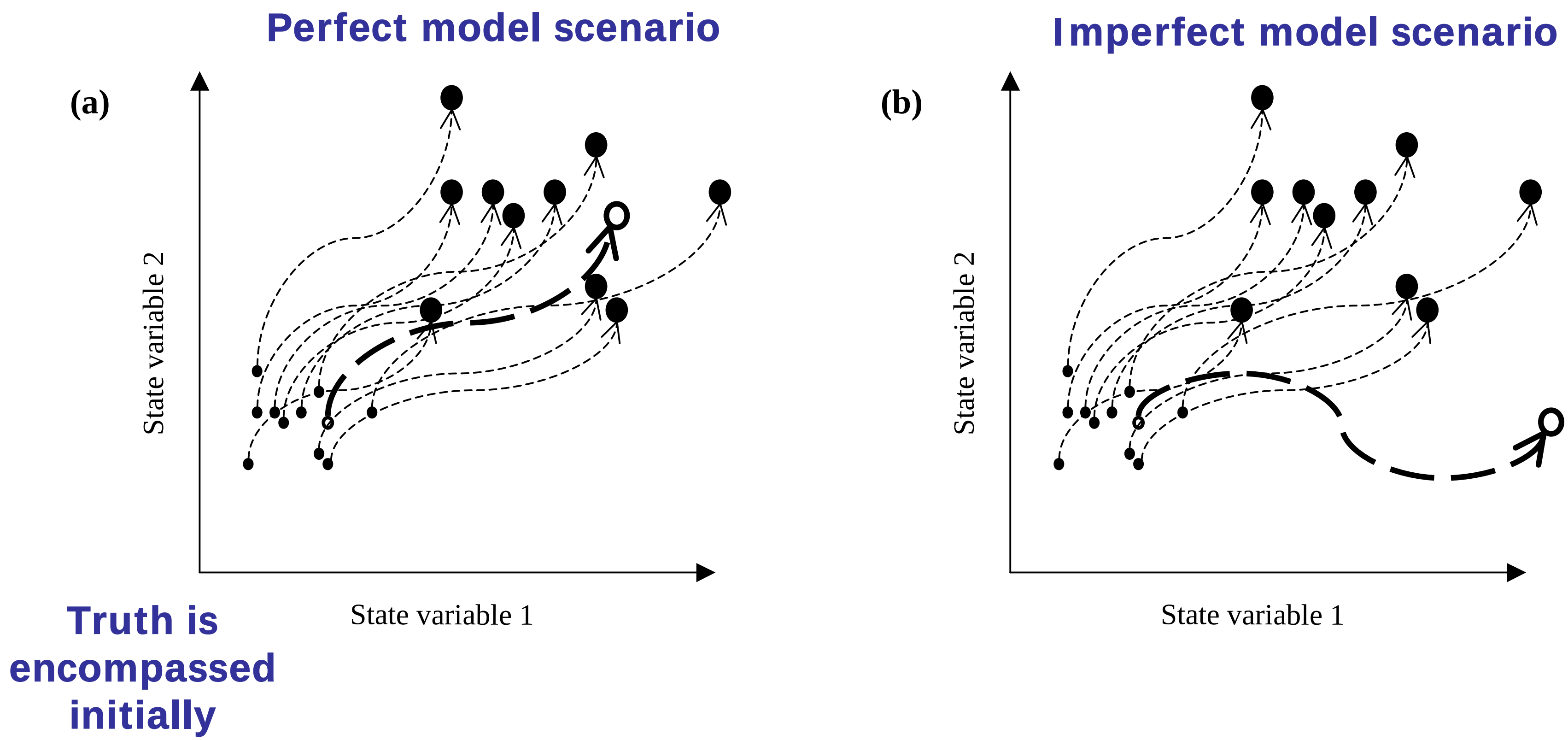
(M = 8-10; Leith 1974)

Spread/Error Correlation

$$\rho^2(\sigma,E) = \frac{2}{\pi}\,\frac{1-\exp(-\beta^2)}{1-\frac{2}{\pi}\exp(-\beta^2)}\ ;\ \beta=\mathrm{std}(\ln \sigma)$$

- Spread/error correlation depends on the time variation of spread
- For constant spread ($\beta=0$) $\rho = 0$.
- Spread is the most useful predictor of skill when it is extreme (large or small)

(Houtekamer 1993; Whitaker and Loughe 1998)



2-D schematic of an idealized ensemble forecast with ten members. **(a)** Perfect model scenario. **(b)** Imperfect model scenario. Initial condition samples are the small, filled dots. Final states are the large dots. Truth is denoted by the unfilled dots and the atmospheric trajectory is the thick, dashed line. Thin, dashed lines are the model trajectories in 2-D phase space.

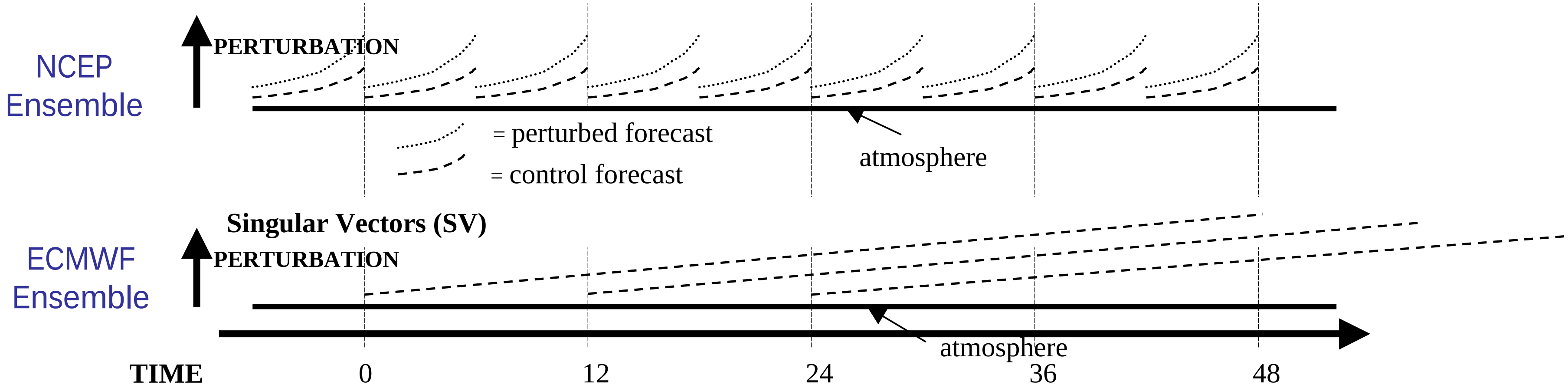
Potential Benefits of Ensemble Forecasting

- Improve deterministic forecast accuracy
  - mean, weighted-mean, best-member, etc.
- Predict forecast skill
  - low (high) spread extremes → high (low) skill
- Define the forecast envelope
  - helps identify potential low probability events → "heads-up" warning for high-impact events
- Identify the limits of predictability
  - point where ensemble spread begins to flat-line
  - point where skill scores become negative
- Make skillful and reliable probability forecasts
  - need to fully define the PDF / possibly using calibration
  - lack of equally likely ICs and presence of model error makes this prospect the most difficult

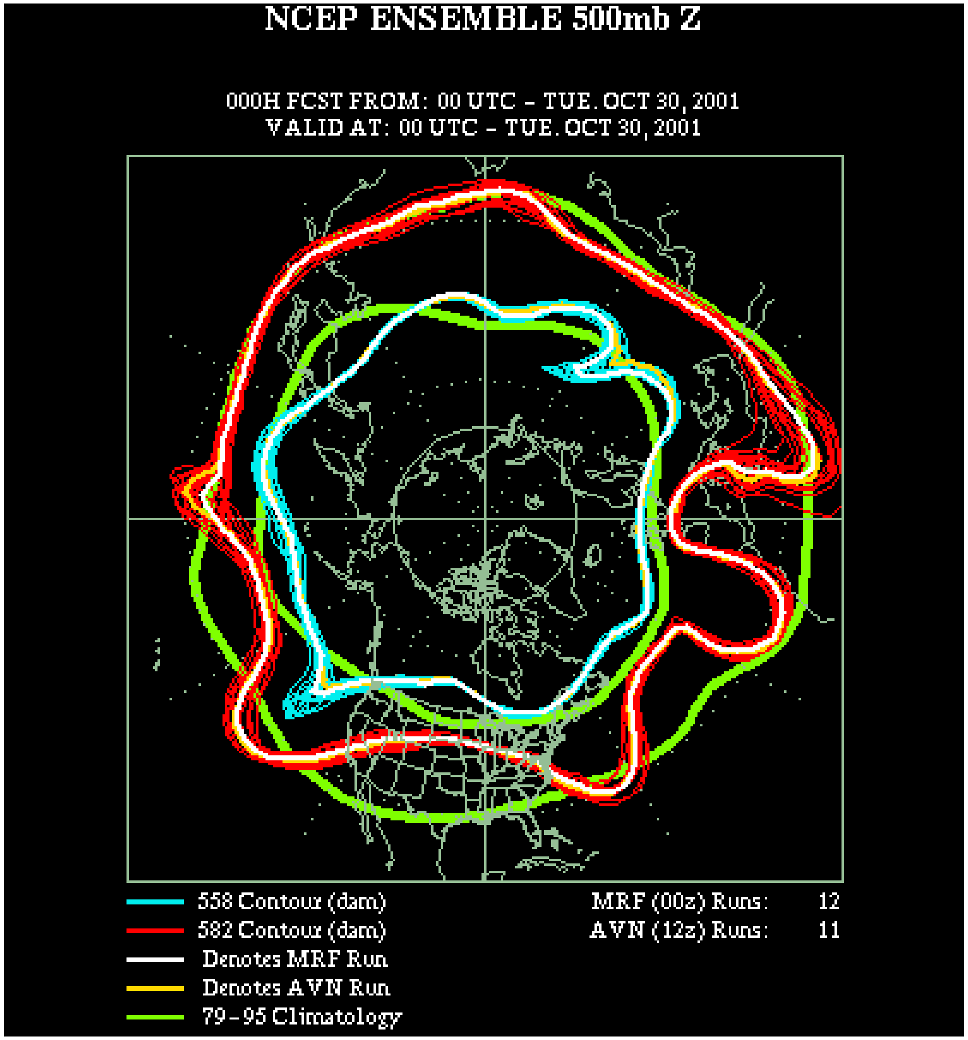
Applications of Ensemble Forecasting

- Initial condition selection strategies
  - Global scale, medium-range (3-14 days)
    - Lagged-average forecasts
      Hoffman and Kalnay (1983)
    - Breeding growing modes (BGM; NCEP)
      Toth and Kalnay (1993, 1997), Tracton and Kalnay (1993)
    - Optimal perturbations, singular vectors (SV; ECMWF)
      Molteni et al. (1996)
    - Random perturbations (Monte Carlo) with dynamic constraints
      Mullen and Baumhefner (1989, 1994)

Breeding Growing Modes (BGM)



NCEP MRF Ensemble



Starts to look a lot like...

Applications of Ensemble Forecasting

- Mesoscale, short-range (0-3 days)
  - Storm and Mesoscale Ensemble Experiment of 1998

    (SAMEX; Hou et al. 2001)

M=25
Δx ~ 30-km {
  - Random perturbations
  - BGM
  - Scaled lagged-average forecasts

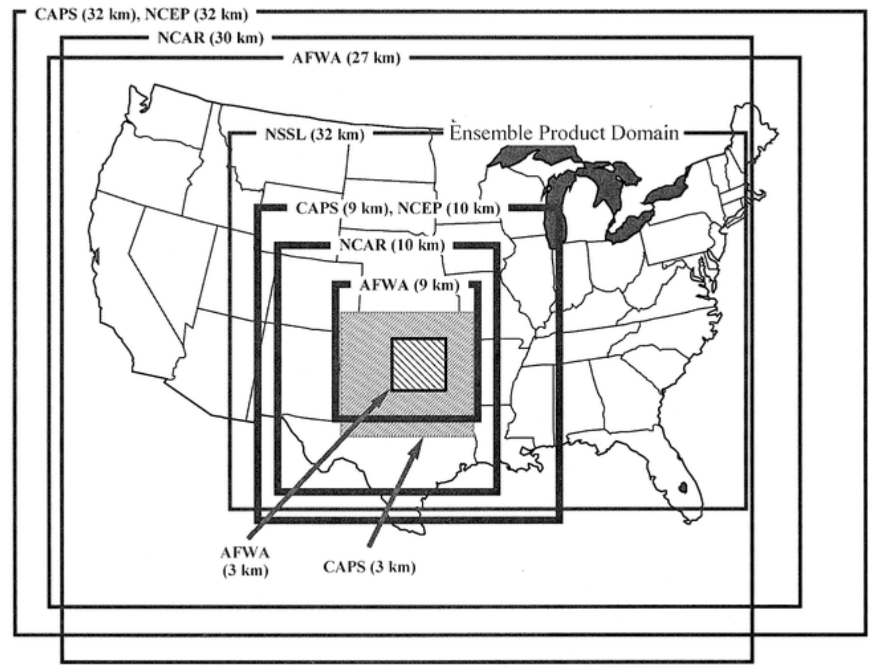
<u>Multi-model, multi-analysis (MMMA) most successful</u>

  - NCEP SREF pilot study / Eta-RSM ensemble

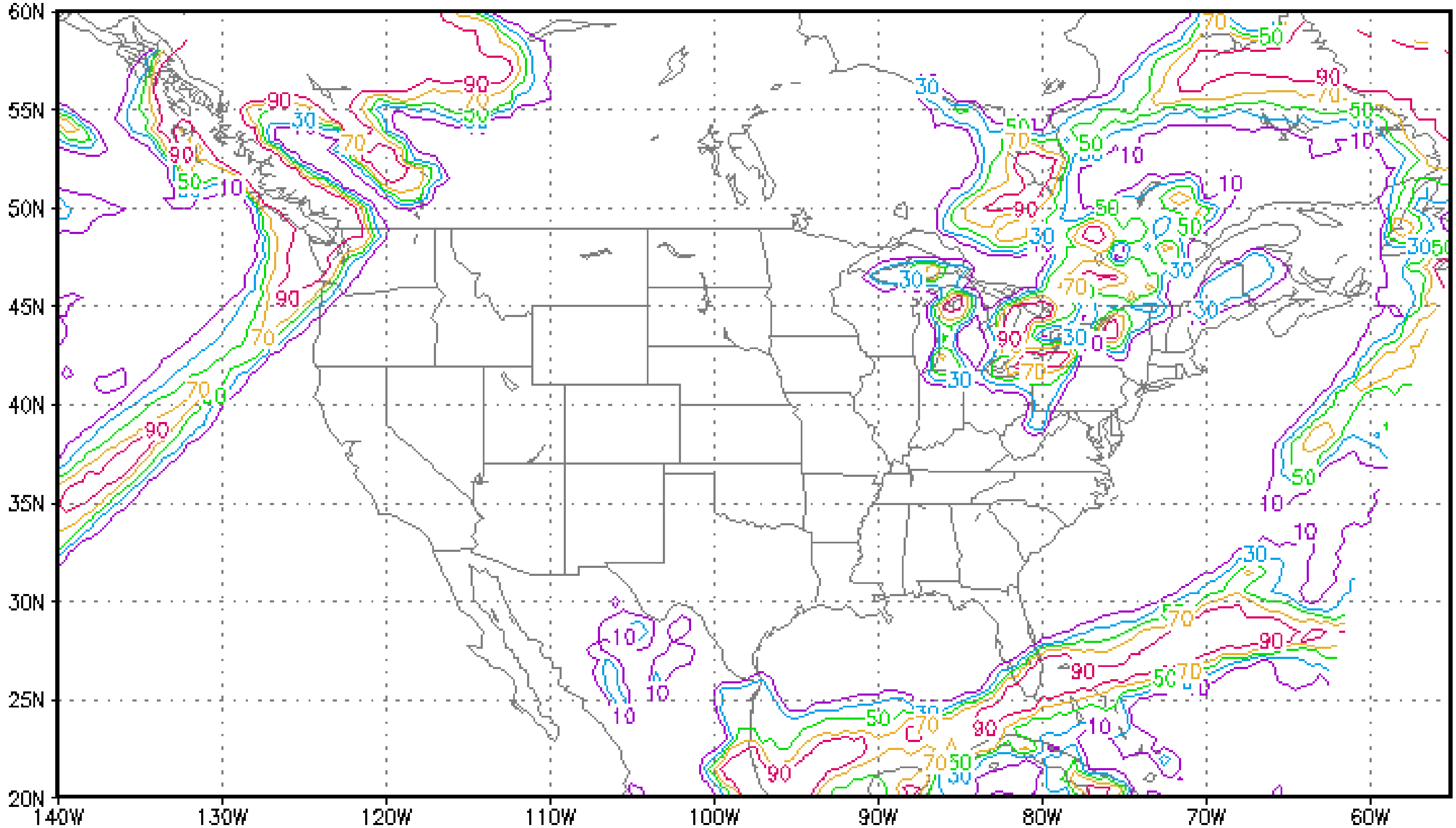
    (Hamill and Colucci 1997, Stensrud et al. 1999)

M=15
Δx ~ 80-km {
  - BGM
  - In-house multi-analysis (MA)

SAMEX '98



COM Prob 12-hr precip > 0.1 in 15H fcst from 09Z 26 OCT 2001
verified time: 00z, 10/27/2001

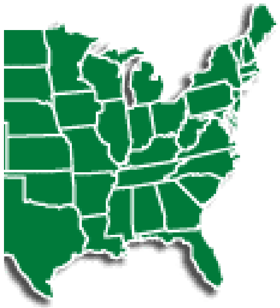


Mesoscale Ensemble Results

- The NCEP Eta-RSM ensemble and SAMEX ensembles matched or outperformed higher-resolution deterministic forecasts by a variety of metrics
  - Ensemble-mean scores better
  - Calibrated Eta-RSM POP forecasts beat NGM MOS
- Spread/error correlations < 0.4, haven't considered extreme spread
- Large missing rates / insufficient spread

Mesoscale Ensemble Results

- Mesoscale, short-range ensemble forecasting (SREF) has been focused primarily over the eastern half of the U.S. where convection plays an important role in atmospheric behavior.
  (Mullen and Baumhefner 1989, 1994; Stensrud and Fritsch 1994a,b; Du et al. 1997; Hamill and Colucci 1997, 1998; Stensrud et al. 2000; Hou et al. 2001)

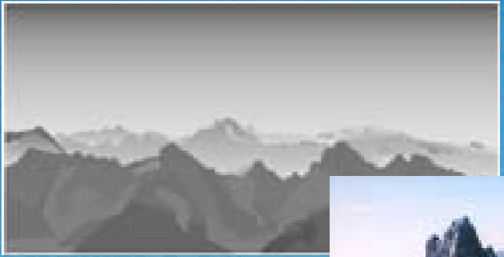


- Error growth due to <u>model deficiencies</u> may be as important as error growth due to imperfect <u>initial conditions</u> (ICs).
- Vary model physics parameterizations or use multiple models to alleviate the under-variability problem.

Outline

- Ensemble Forecasting Background
- **Motivation for this Research**
- The Approach
- Results
- Conclusions
- Future Work

Local (Pac NW) Factors



- Convection is weaker, shallower, and less frequent – lower-order chaotic effect.
- Mesoscale structures are determined predominately by the interaction of the synoptic-scale flow with the regional orography.
- Vast data sparse region over Pacific can lead to large phase and amplitude errors in synoptic-scale features.

- Model deficiencies may be less important than over the eastern half of the U.S.
- Therefore, in the Pacific Northwest, imperfect ICs are the primary concern.

Motivation

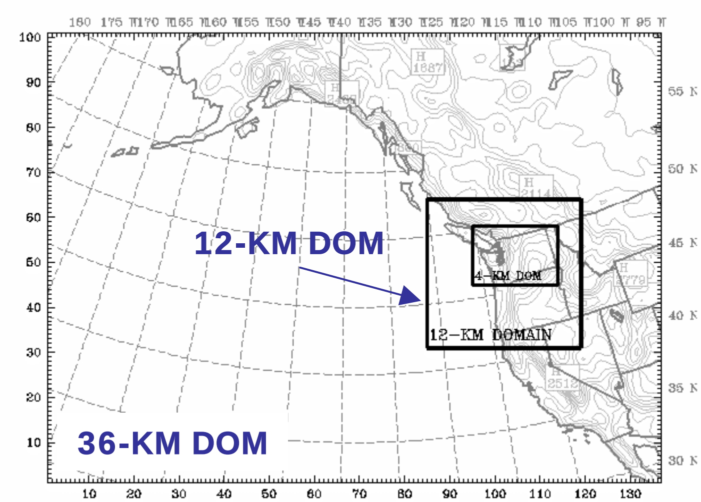
- Diminishing returns experienced by increasing horizontal grid resolution
- Comprehensively test SREF in the Pacific Northwest
- Test the multi-analysis approach in earnest
  - Using less correlated analyses
  - Using a large number of cases

Outline

- Ensemble Forecasting Background
- Motivation for this Research
- **The Approach**
- Results
- Conclusions
- Future Work

The Approach

- Single modeling system (MM5)
- 2-day (48-hr) forecasts at 36-km & 12-km grid spacing in real-time.



- IC Selection: Multi-analysis (MA)

  [From different operational centers (NCEP, CMC, FNMOC)]

  David Richardson (QJRMS 2001): An ensemble using this MA strategy can realize up to 80% of the improvements gained by running a multi-model, multi-analysis (MMMA) ensemble.
- Lateral Boundary Conditions (LBCs): Drawn from the corresponding, synoptic-scale forecasts

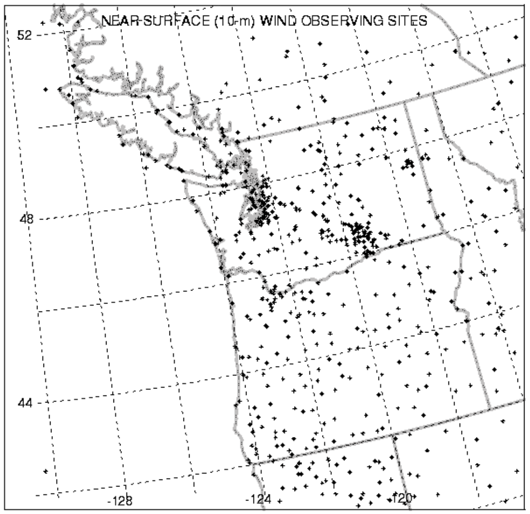
Verification Method

- Model forecasts are interpolated to the observation sites over 12-km domain only.
- Focus is on near-surface weather variables.
- Mesoscale verification variable of choice:

10-m Wind Direction

  - More extensive coverage & greater # of reporting sites.
  - Greatly influenced by regional orography and synoptic scale changes.
  - MM5's systematic biases in the other near-surface variables can dominate errors originating from ICs.
- Probability forecast verification variable:

12-h Accumulated Precipitation



Error

Mean Absolute Error (MAE)

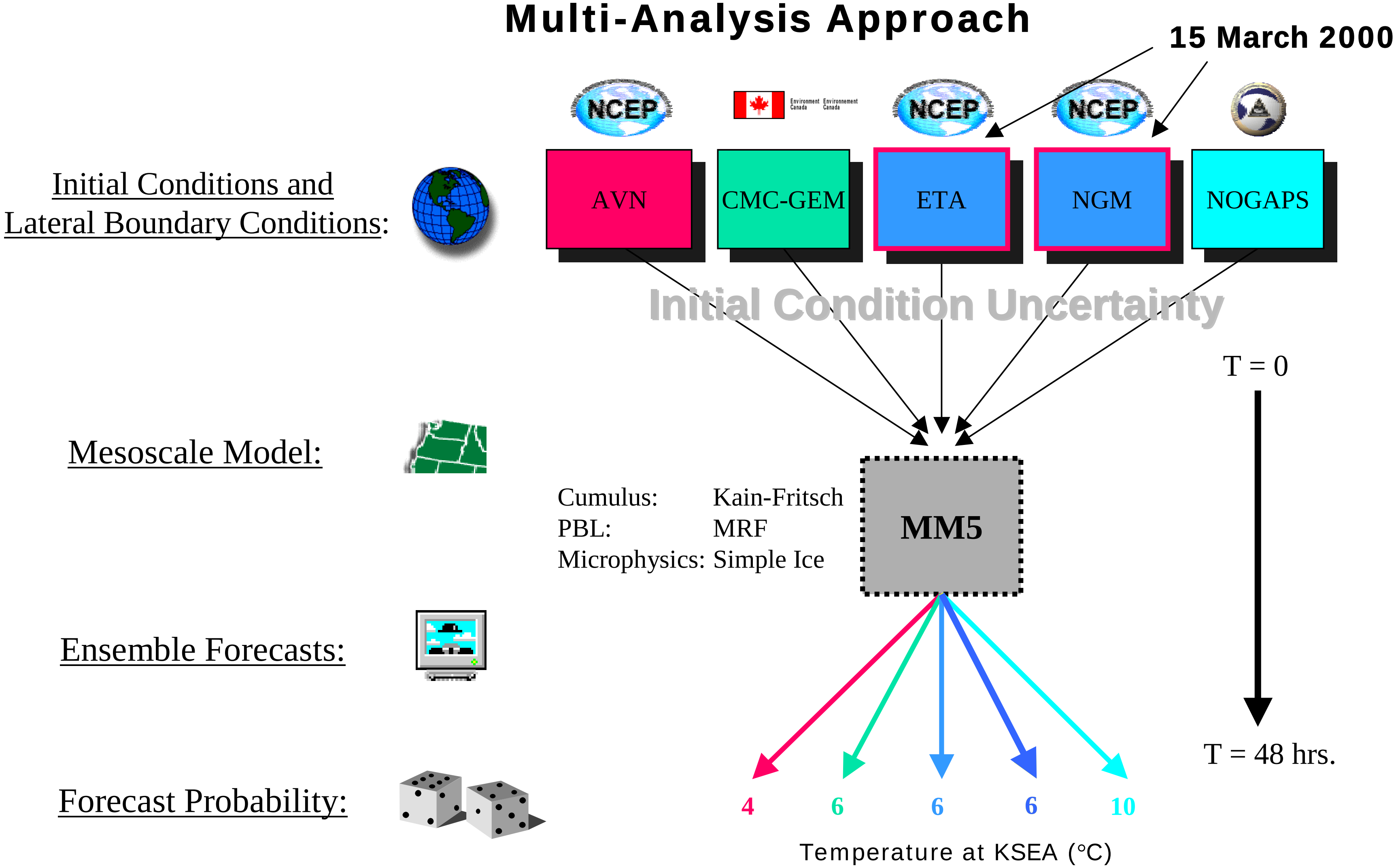
Spread

Dom. Av. Standard Dev. (σ)

PHASE I: JAN - JUN 2000

N=102

Multi-Analysis Approach

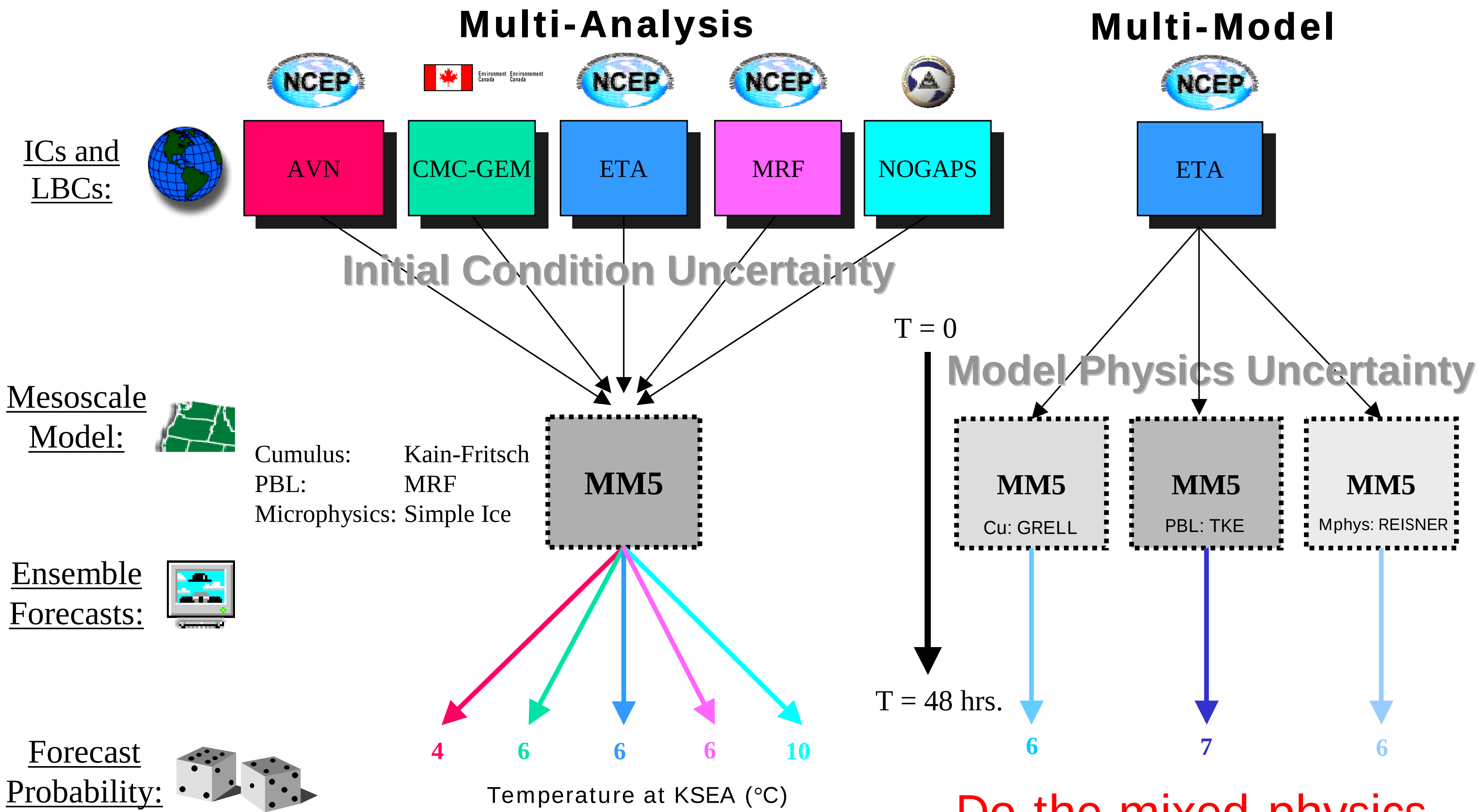


$\bar{x}$ = 6.4 °C   s = 2.2 °C

P(t > 6.4 °C) = 50% **???**

PHASE II: OCT 2000 - MAR 2001

N=114



$\bar{x}$ = 6.4 °C   s = 2.2 °C

P(t > 6.4 °C) = 50% **???**

Do the mixed-physics members add useful variability?

Limitations

- ICs are not equally likely nor equally skillful.
- Only a finite number of analyses available.
- MA approach relies on products that may evolve with time.
  - Examples:
    - 15 March 2000 changes to NGM (RDAS -> EDAS)
    - Fall 2000 changes to ETA (32-km -> 22-km, etc.)
- Complete independence between the IC "perturbations" (the differences between the analyses) is not guaranteed, thus the ICs may not be a true random sample.

Benefits

- Avoids potentially unrealistic random perturbations and bred modes.
- Favors a more realistic set of initial conditions, that may also better approximate the analysis uncertainty.
- Has the largest initial spread of any method.
- Low computational expense – do not have to calculate bred modes or singular vectors.

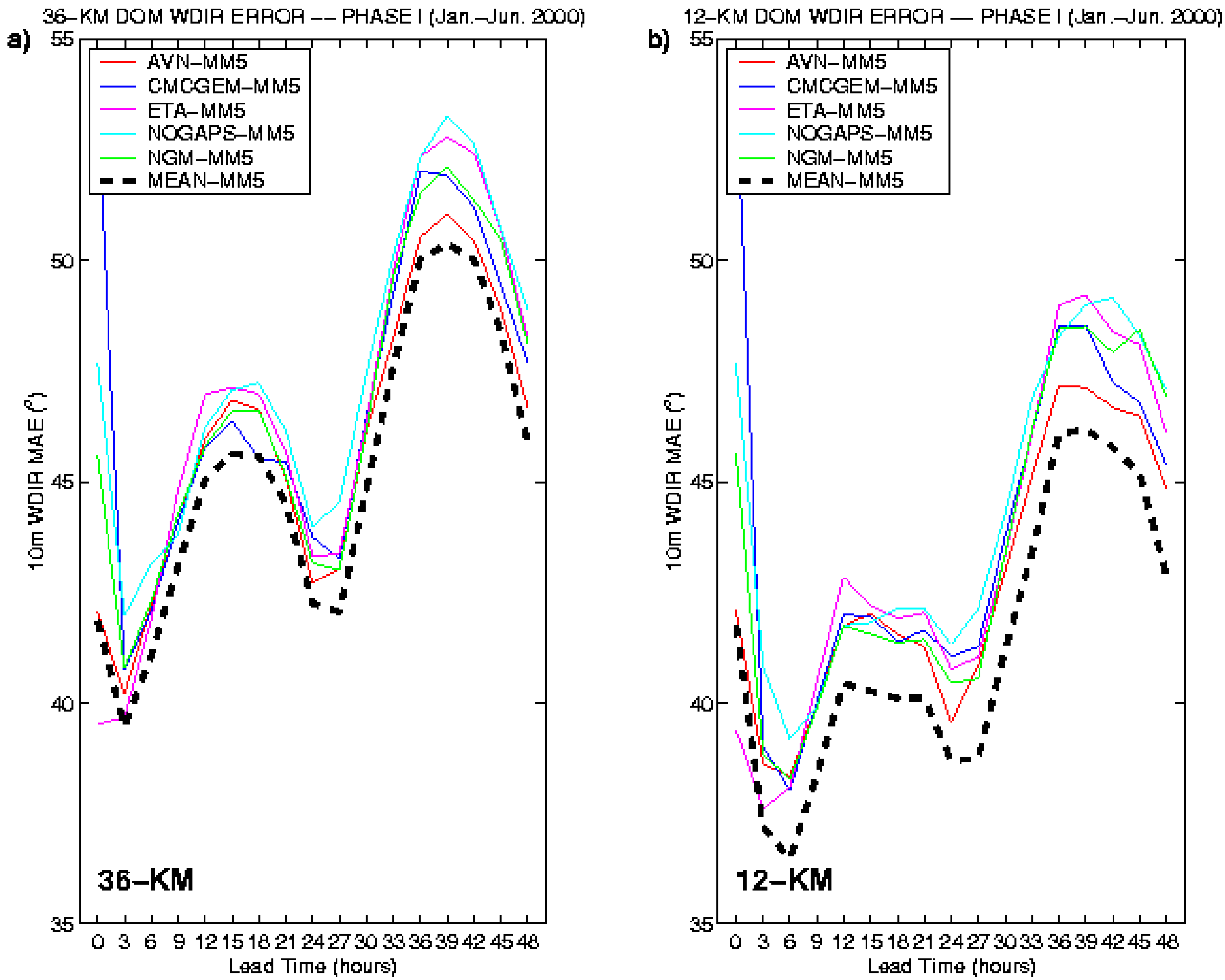
Research Questions

1) Is a MA ensemble approach using ICs and LBCs from different operational forecast systems viable?
2) Does the ensemble-mean possess greater skill than its component forecasts in terms of standard measures of forecast skill?
3) How does ensemble-mean forecast skill compare with higher-resolution deterministic forecasts?
4) Can the mesoscale ensemble predict forecast skill? Specifically, is there a significant correlation between ensemble spread and ensemble-mean skill and/or the skill of the component forecasts?
5) How accurate is the ensemble-forecasted probability distribution? Does the verification lie outside the ensemble envelope more often than it should? How reliable are the probability forecasts?
6) What are the effects of adding physics diversity to the ensemble-mean skill, the spread/error correlations and the distribution of the forecasts?

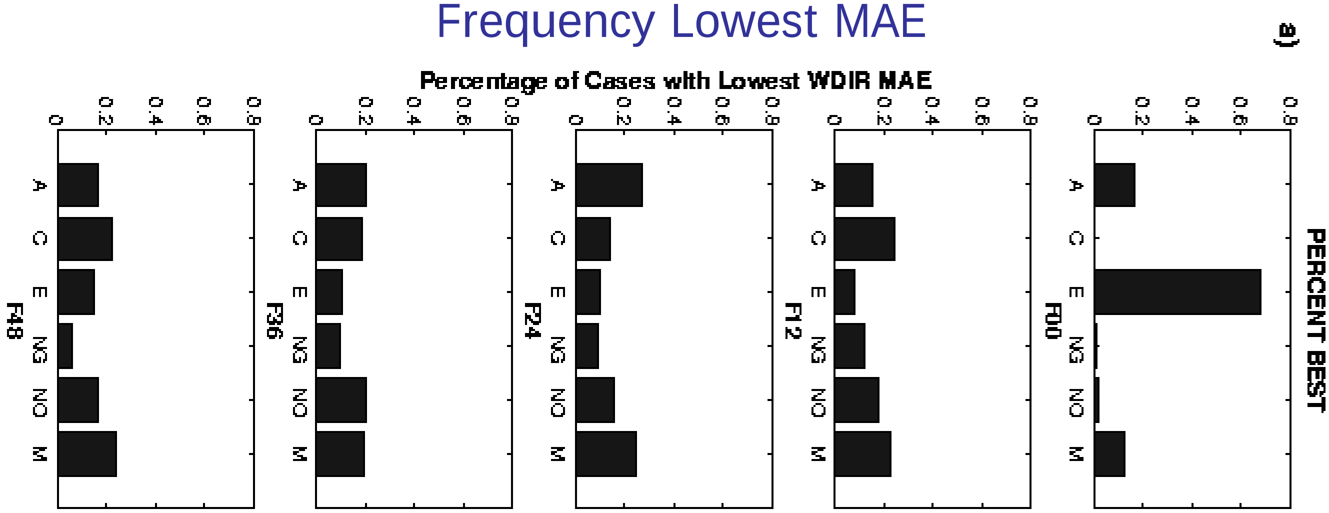
Outline

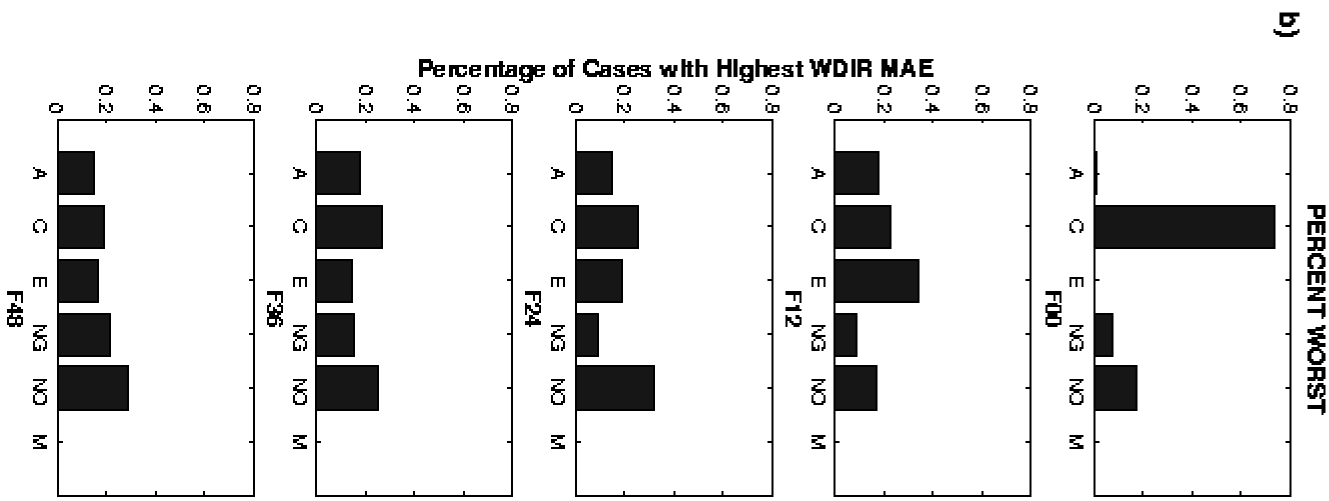
- Ensemble Forecasting Background
- Motivation for this Research
- The Approach
- **Results**
  - **Ensemble-mean skill**
  - **Forecast skill prediction**
  - **Ensemble forecast distribution**
  - **The effects of physics diversity**
- Conclusions
- Future Work

10-m Wind Direction Errors (MAE)



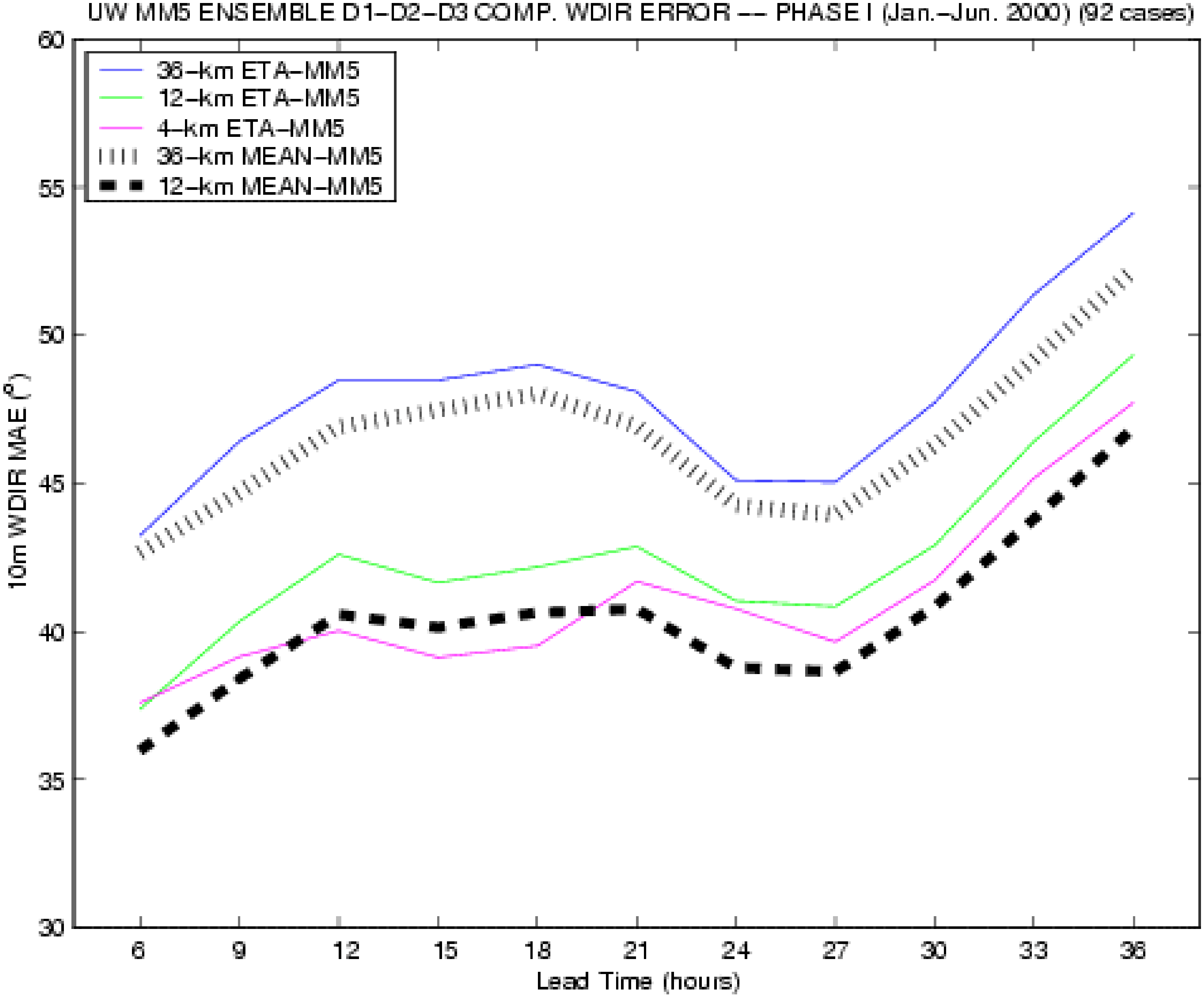
Frequency Lowest MAE



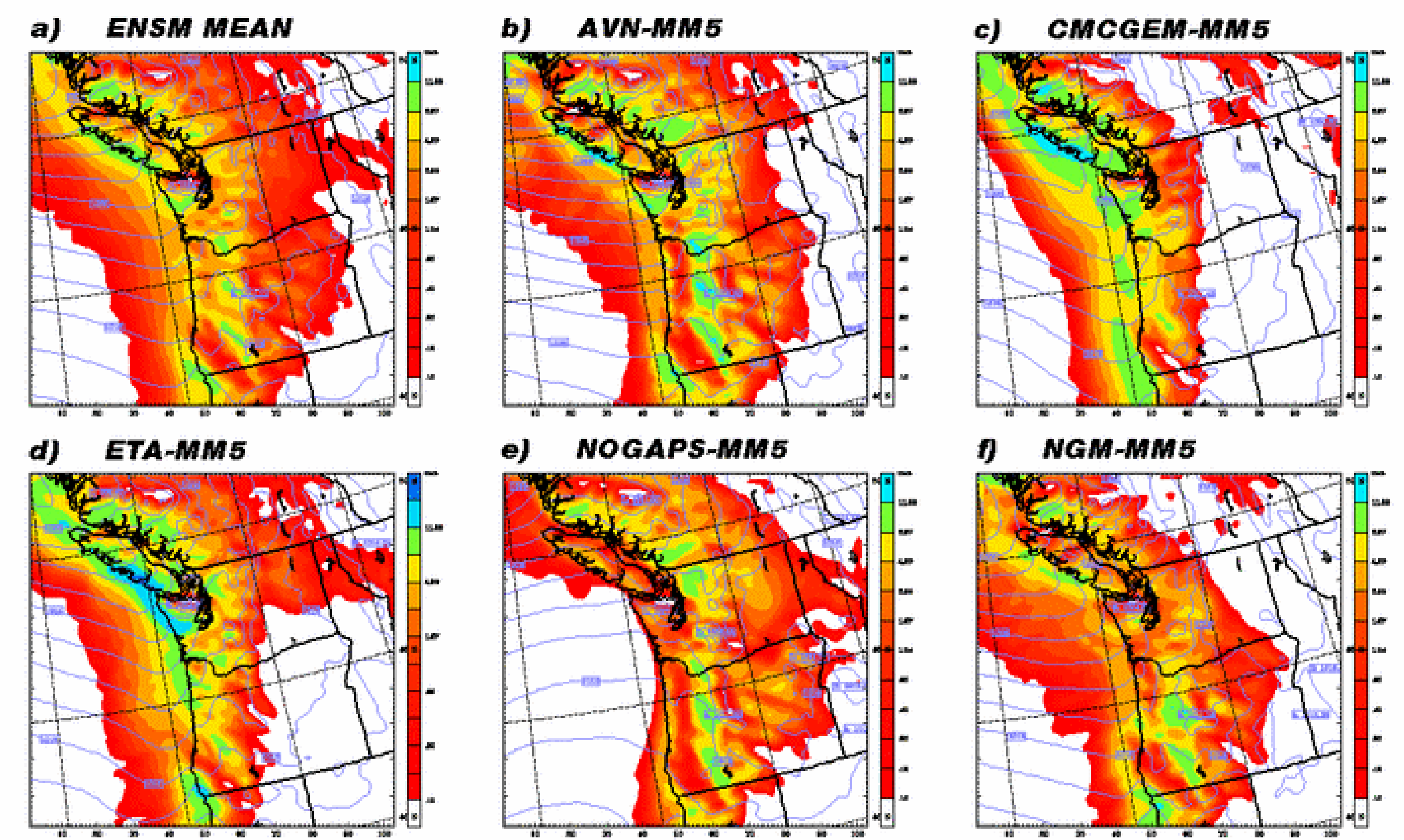


Frequency Highest MAE

Ensemble Averaging vs. Increasing Grid Resolution

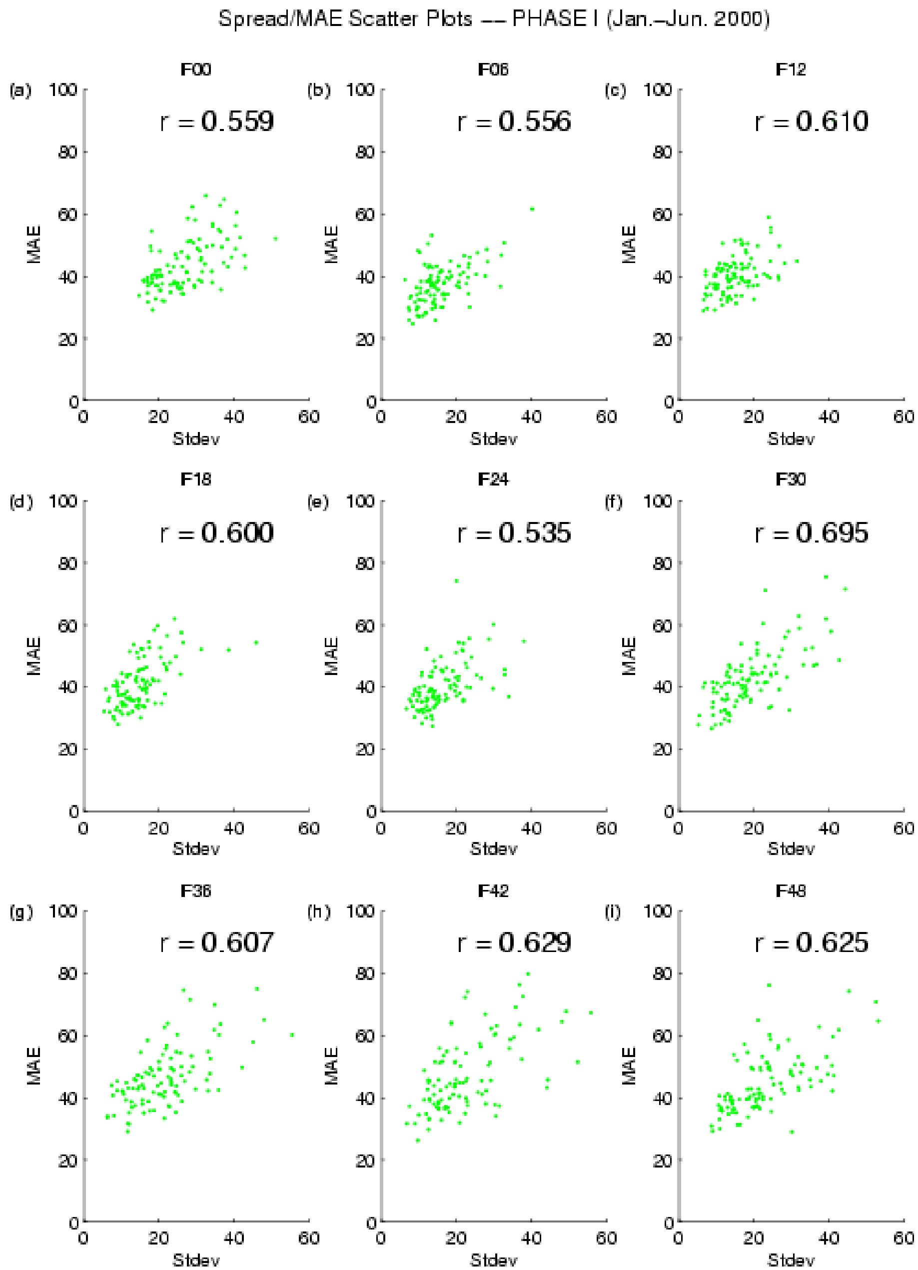


3 March 2000 / 0000 UTC / PCP_3 F30

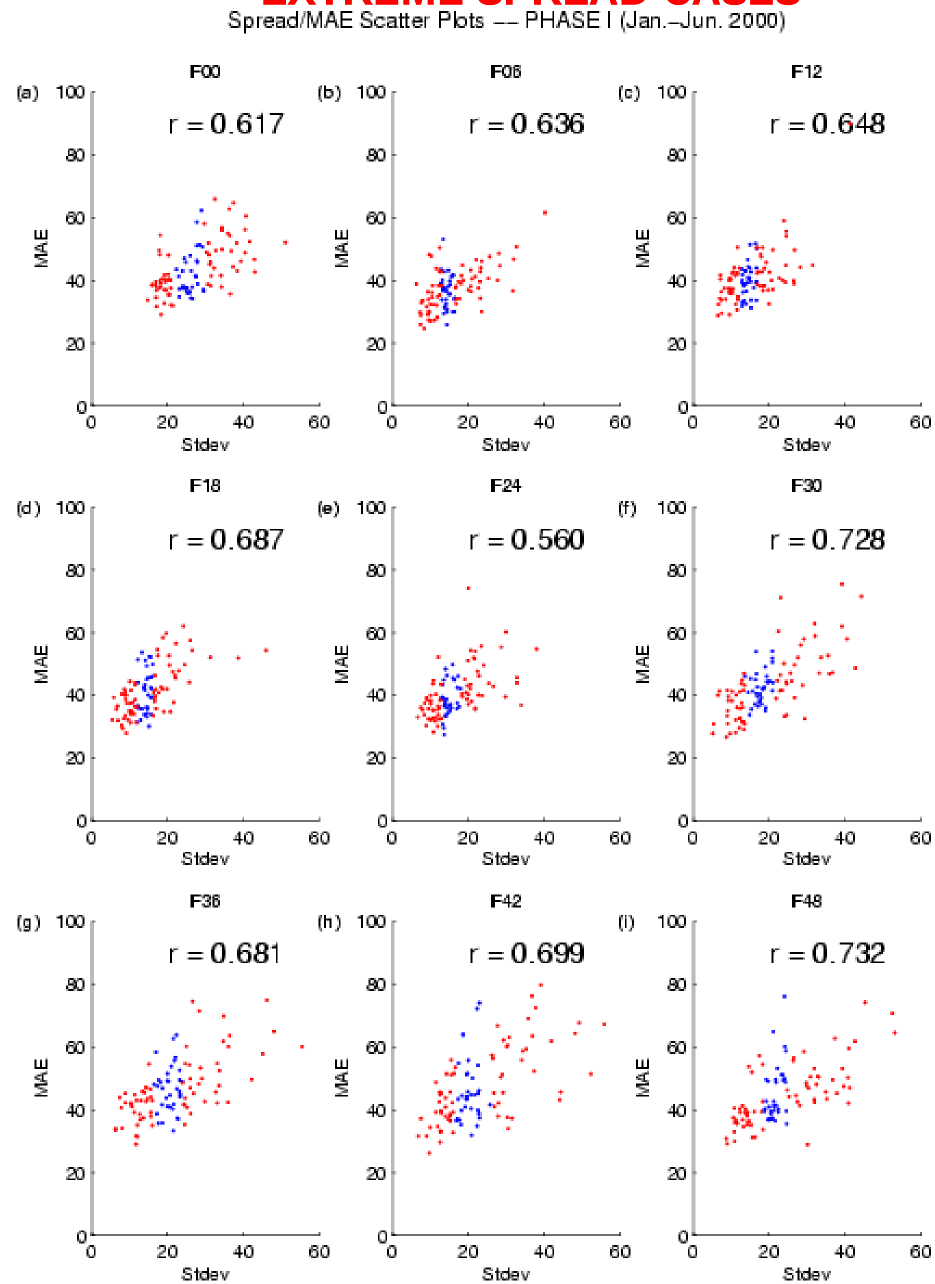


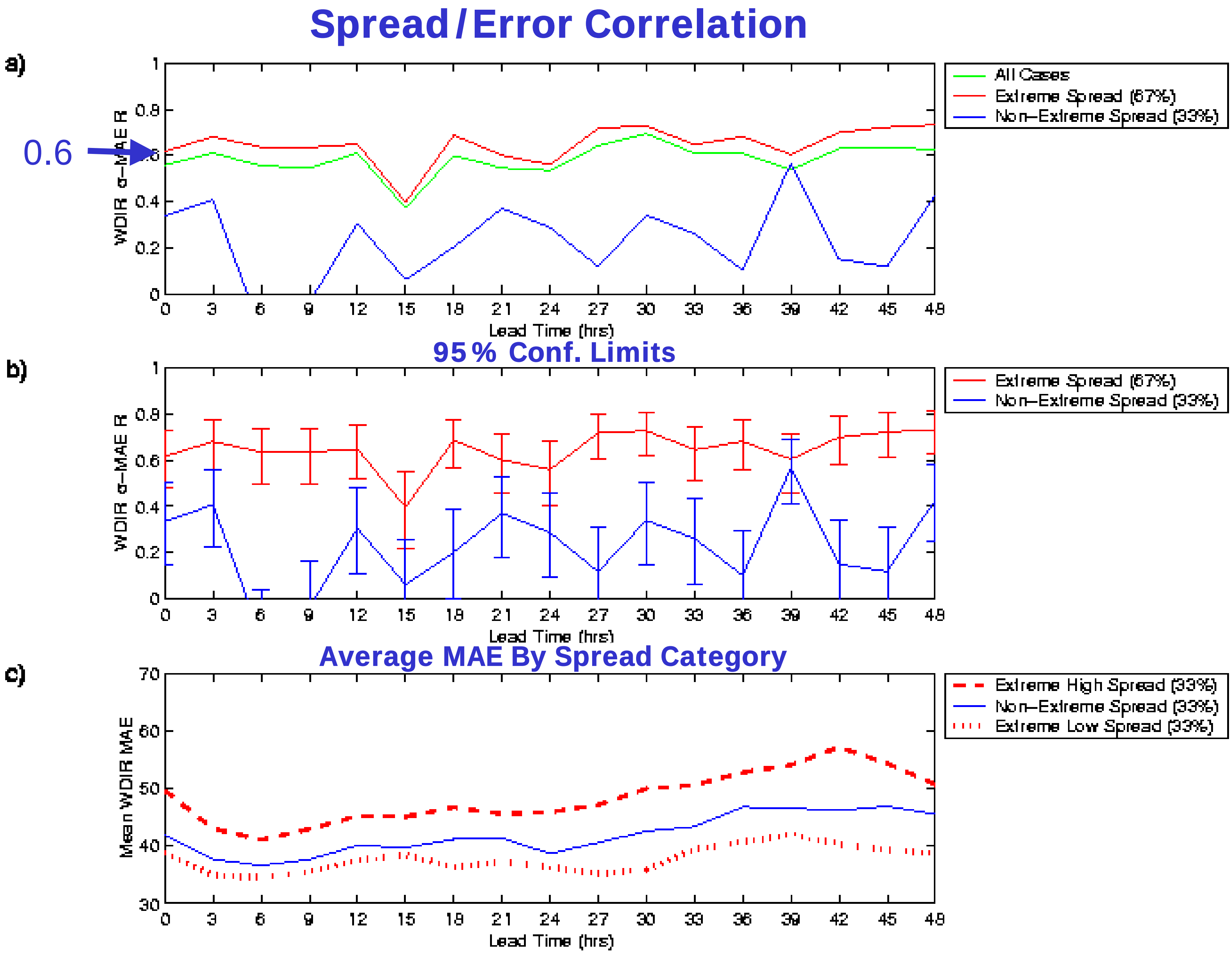
Spread/Error Scatter Diagrams

ALL CASES

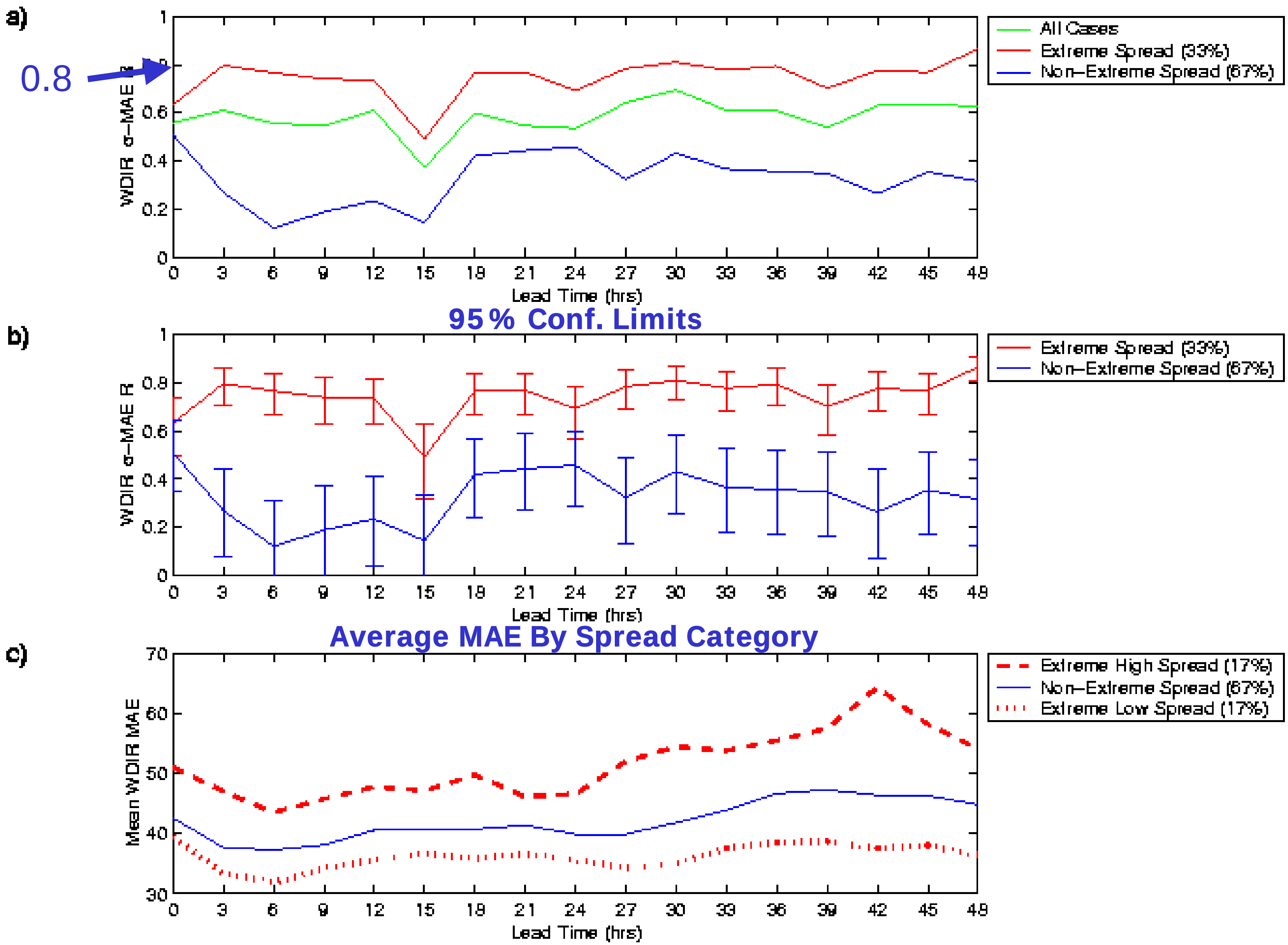


EXTREME SPREAD CASES

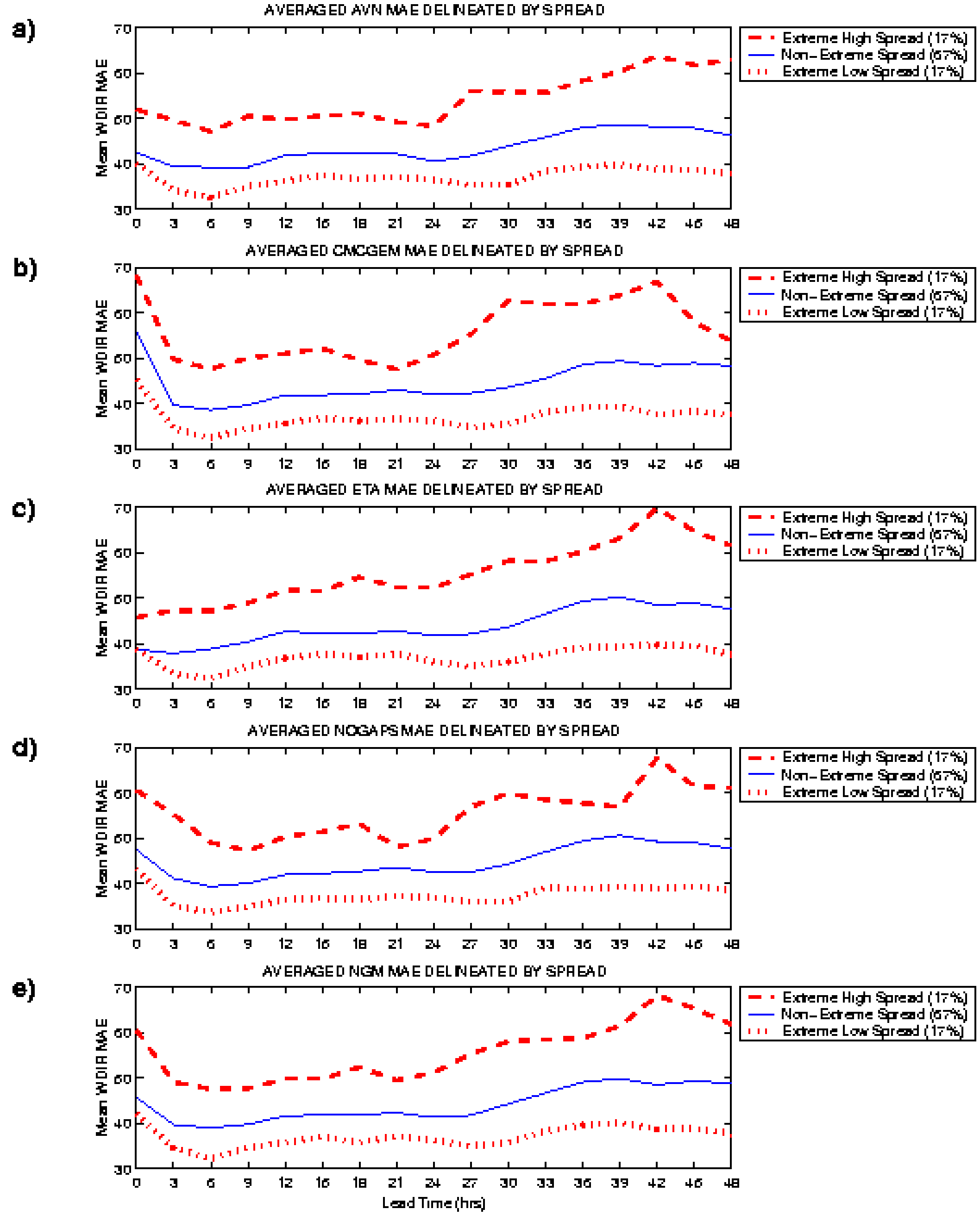



Spread/Error Correlation
a)
0.6
WDIR σ–MAE R
Lead Time (hrs)
All Cases
Extreme Spread (67%)
Non–Extreme Spread (33%)
95% Conf. Limits
b)
WDIR σ–MAE R
Lead Time (hrs)
Extreme Spread (67%)
Non–Extreme Spread (33%)
Average MAE By Spread Category
c)
Mean WDIR MAE
Lead Time (hrs)
Extreme High Spread (33%)
Non–Extreme Spread (33%)
Extreme Low Spread (33%)

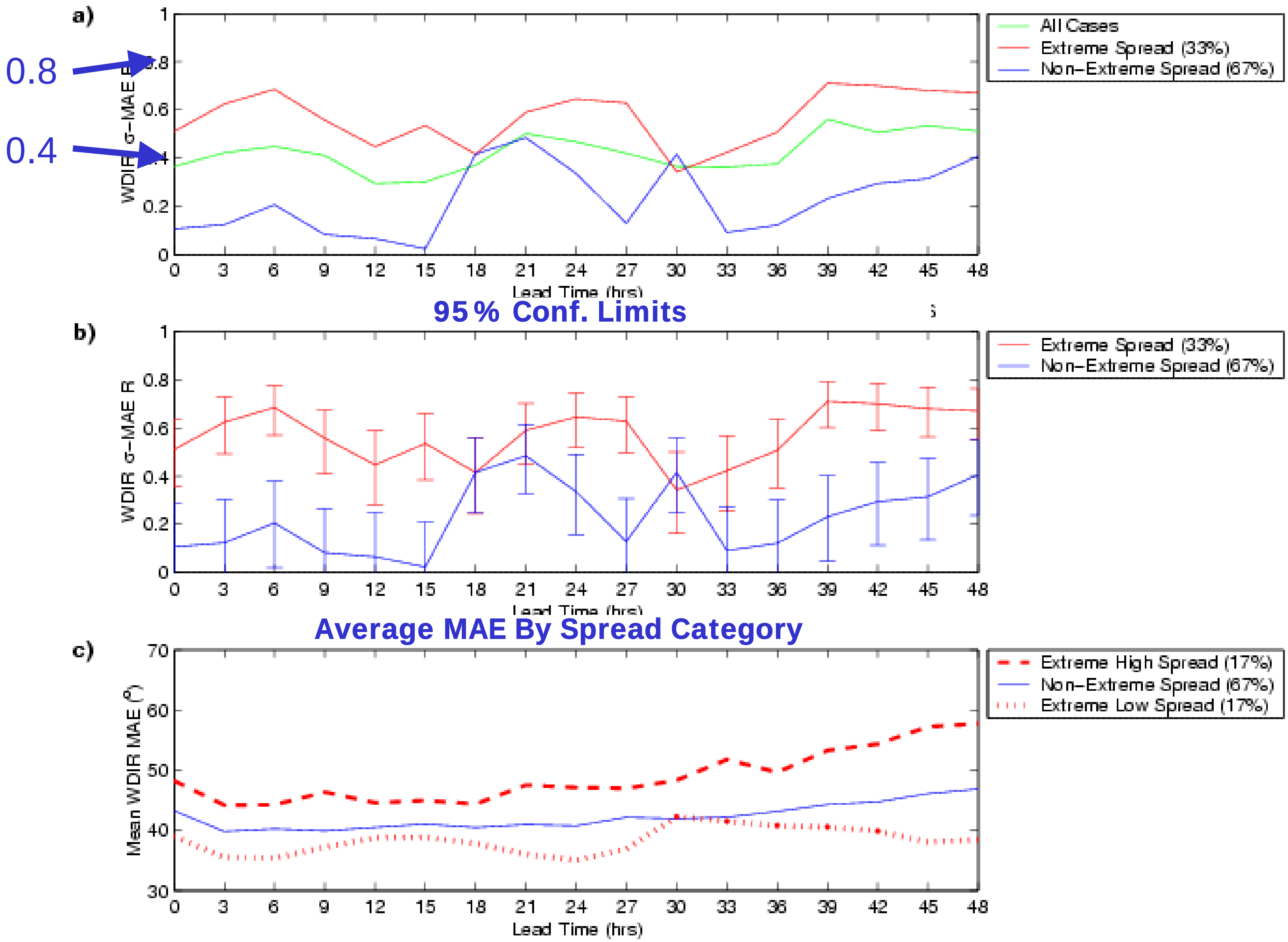
Spread / Error Correlation



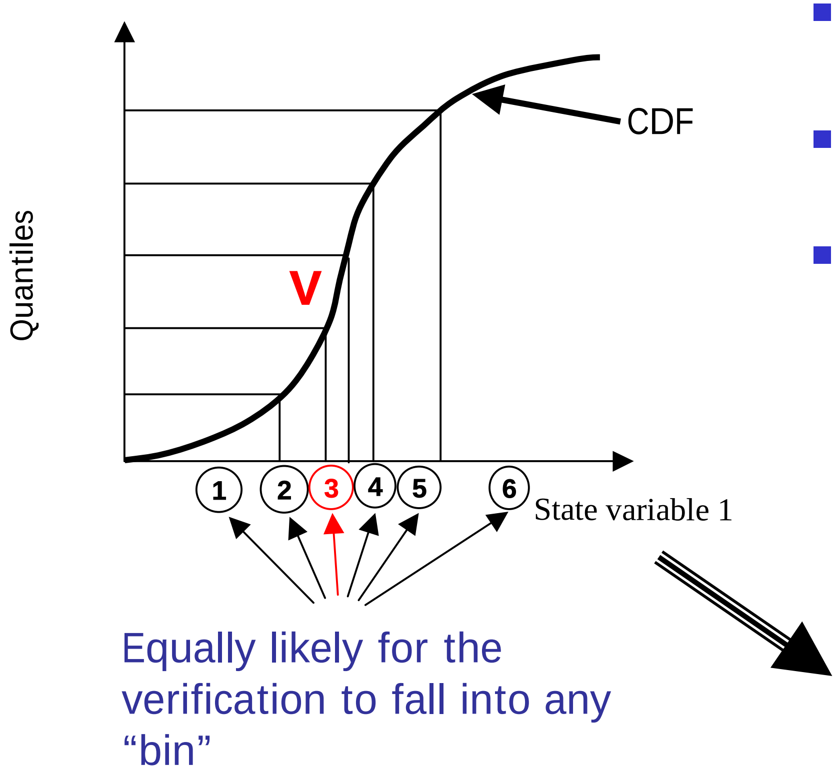
Average MAE By Spread Category



Spread/Error Correlation

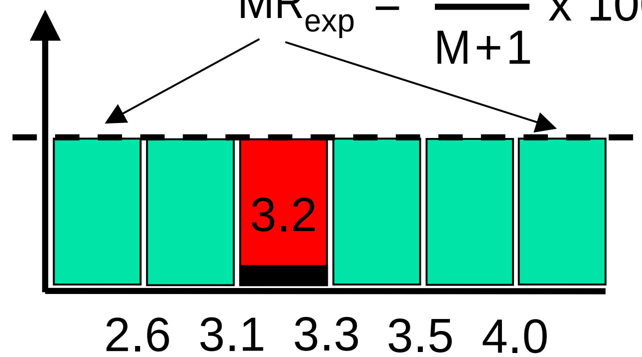


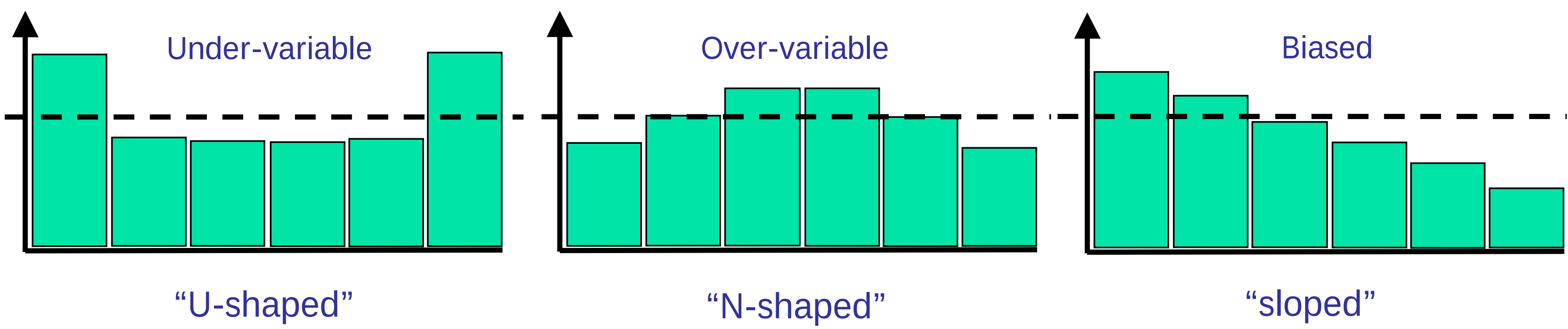
The Verification Rank Histogram



- ICs must be random & <u>equally likely</u> [independent, identically distributed (iid)]
- The verification should appear no different than any other IC
- It is possible that the verifying value falls outside, but still appears as a plausible member of the ensemble

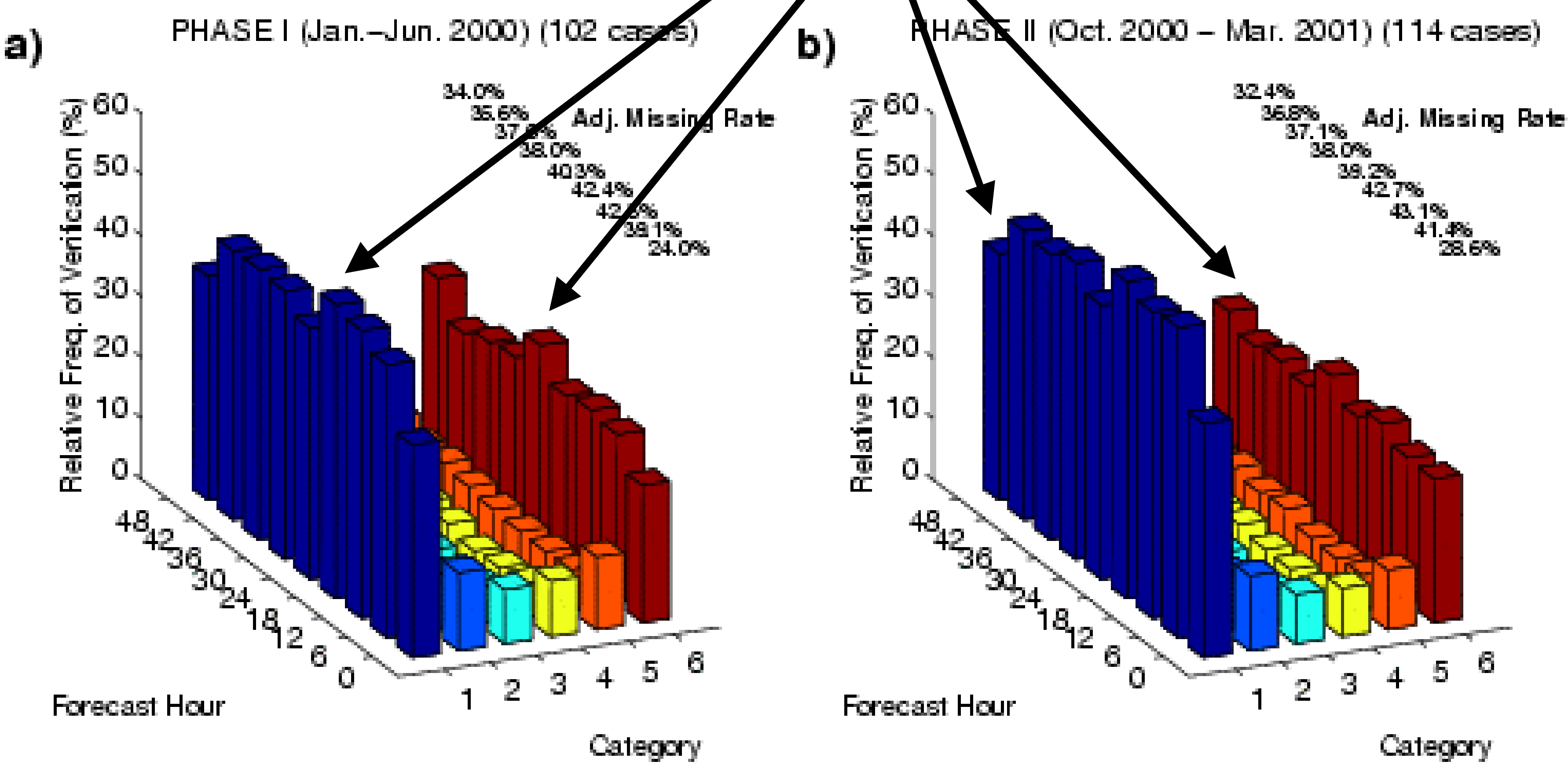
$$MR_{exp} = \frac{2}{M+1} \text{ x } 100\% = 33\%$$



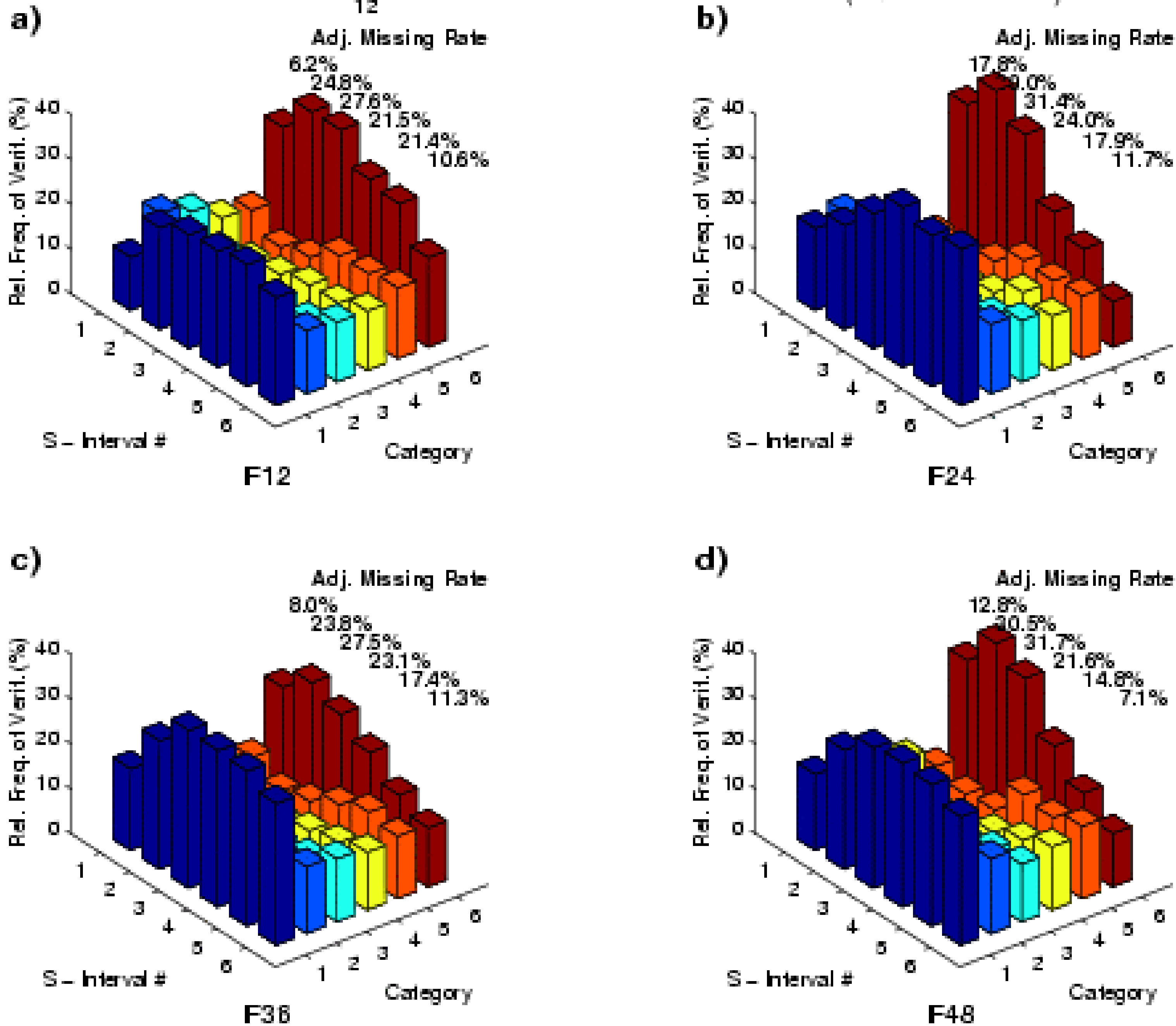


Missing rates
WAY too large!

$$>> \frac{2}{M+1}$$



12–KM PCP$_{12}$ RANK HISTOGRAMS –– PHASE I (Jan.–Jun. 2000)



Evaluating probability forecasts

Brier Score (BS):

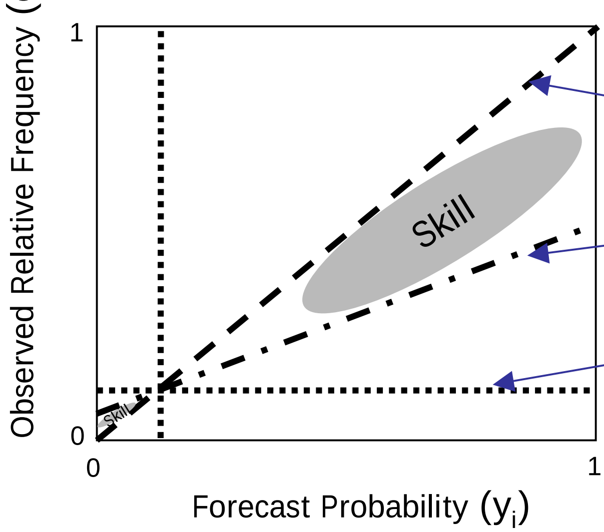
$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

"Mean-square error of probability forecasts"

$o_i = \{$ 1 occurrence
0 non-occurrence

N = number of forecast/event pairs
M = number of ensemble members
(M+1 is the # of probability categories)

Reliability Diagram



BS = REL - RES + UNC

Better: ↓ ↓ ↑ ↓

REL = reliability (conditional squared bias)

$$REL = \frac{1}{N}\sum_{i=1}^{M+1} N_i (y_i - \bar{o}_i)^2$$

RES = resolution (event discriminating ability)

$$RES = \frac{1}{N}\sum_{i=1}^{M+1} N_i (\bar{o}_i - \bar{o})^2 \qquad \bar{o} = \sum_{i=1}^{M+1} \bar{o}_i$$

$0 \leftarrow p_i \rightarrow 1$

UNC = uncertainty (depends only on obs)

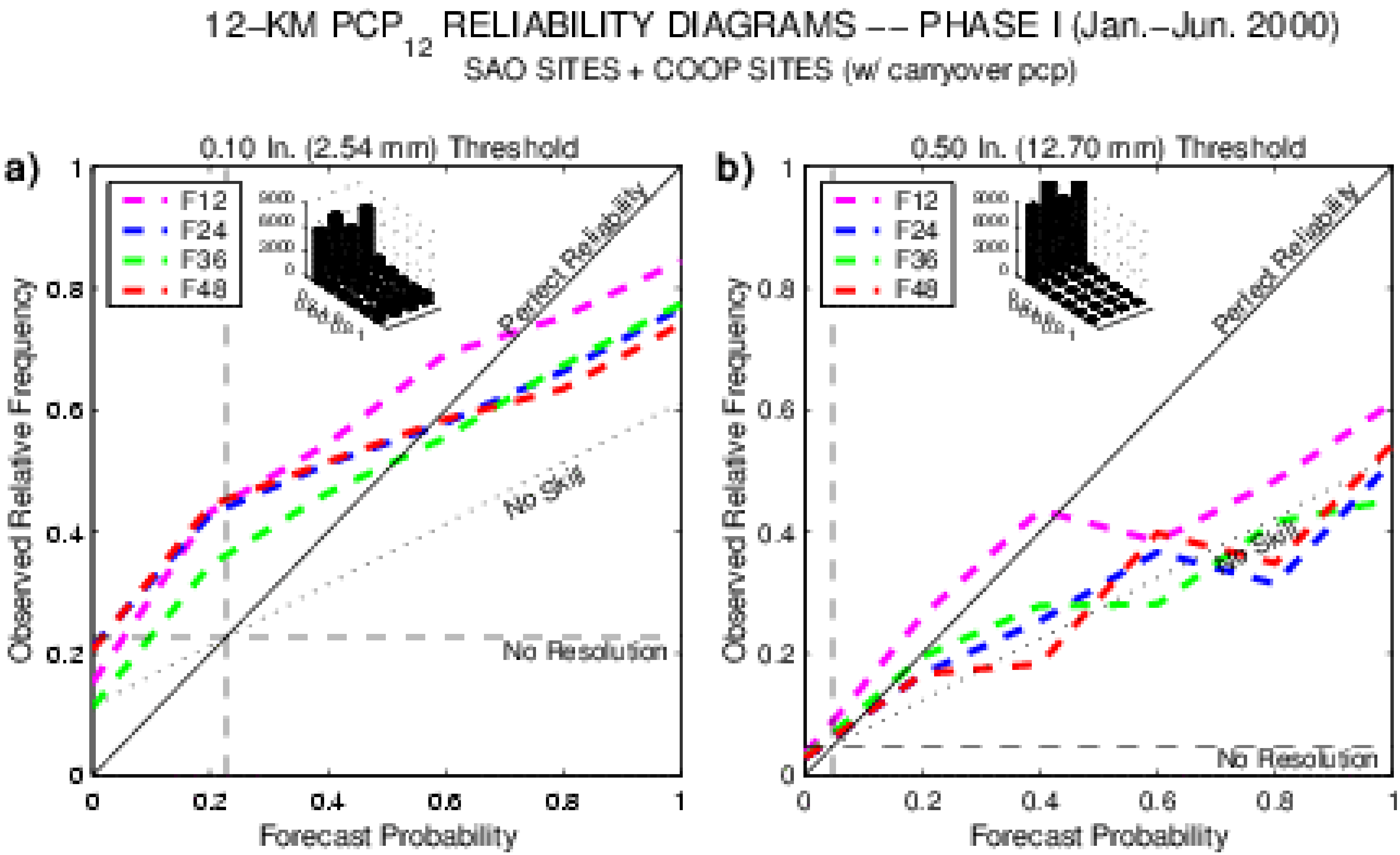
$$UNC = \bar{o}(1 - \bar{o})$$

(1:1) line – Perfect reliability

"No Skill" line (REL = RES)

"No Resolution" line ($\bar{o}_i$ = UNC)

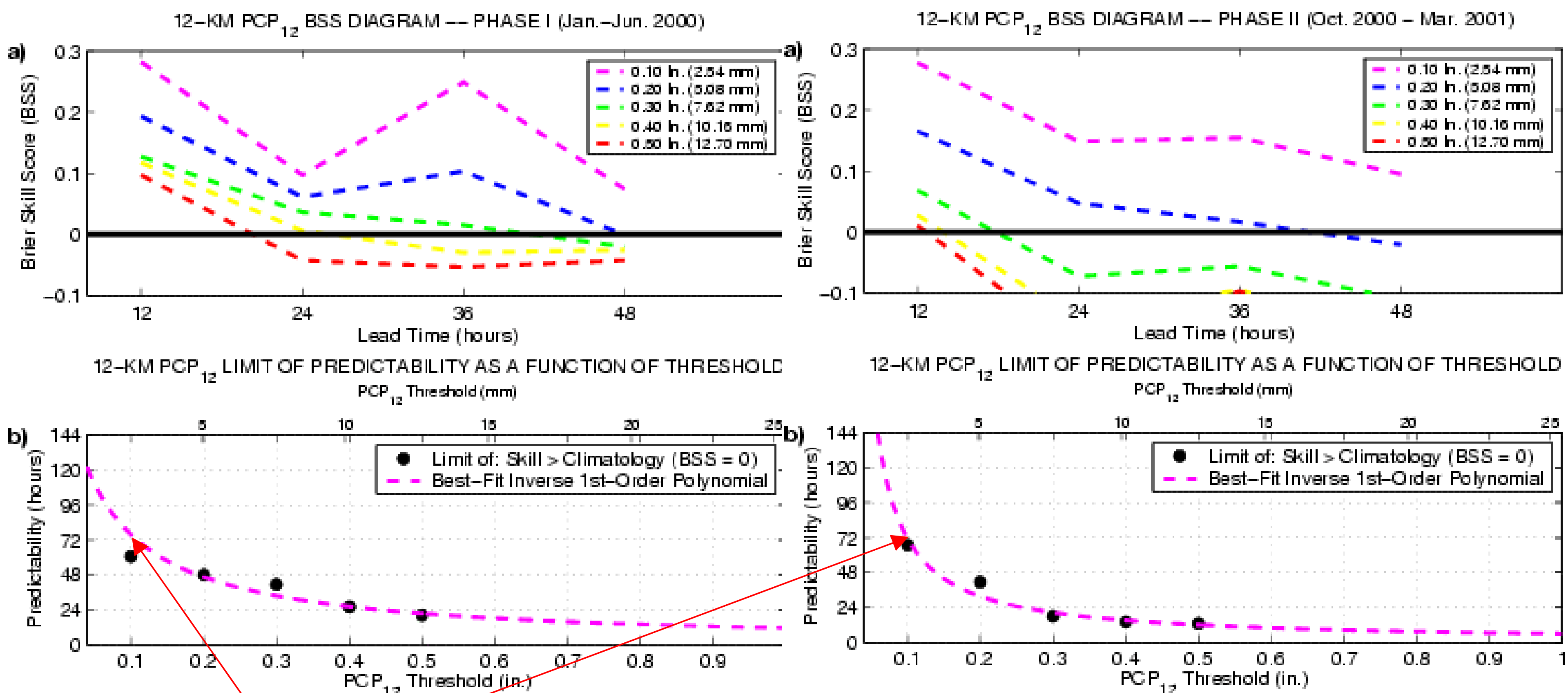
Reliability Diagrams



Brier Skill Score:

$$BSS = \frac{BS_{cli} - BS}{BS_{cli}} = \frac{RES - REL}{UNC} \quad (BS_{cli} = UNC)$$

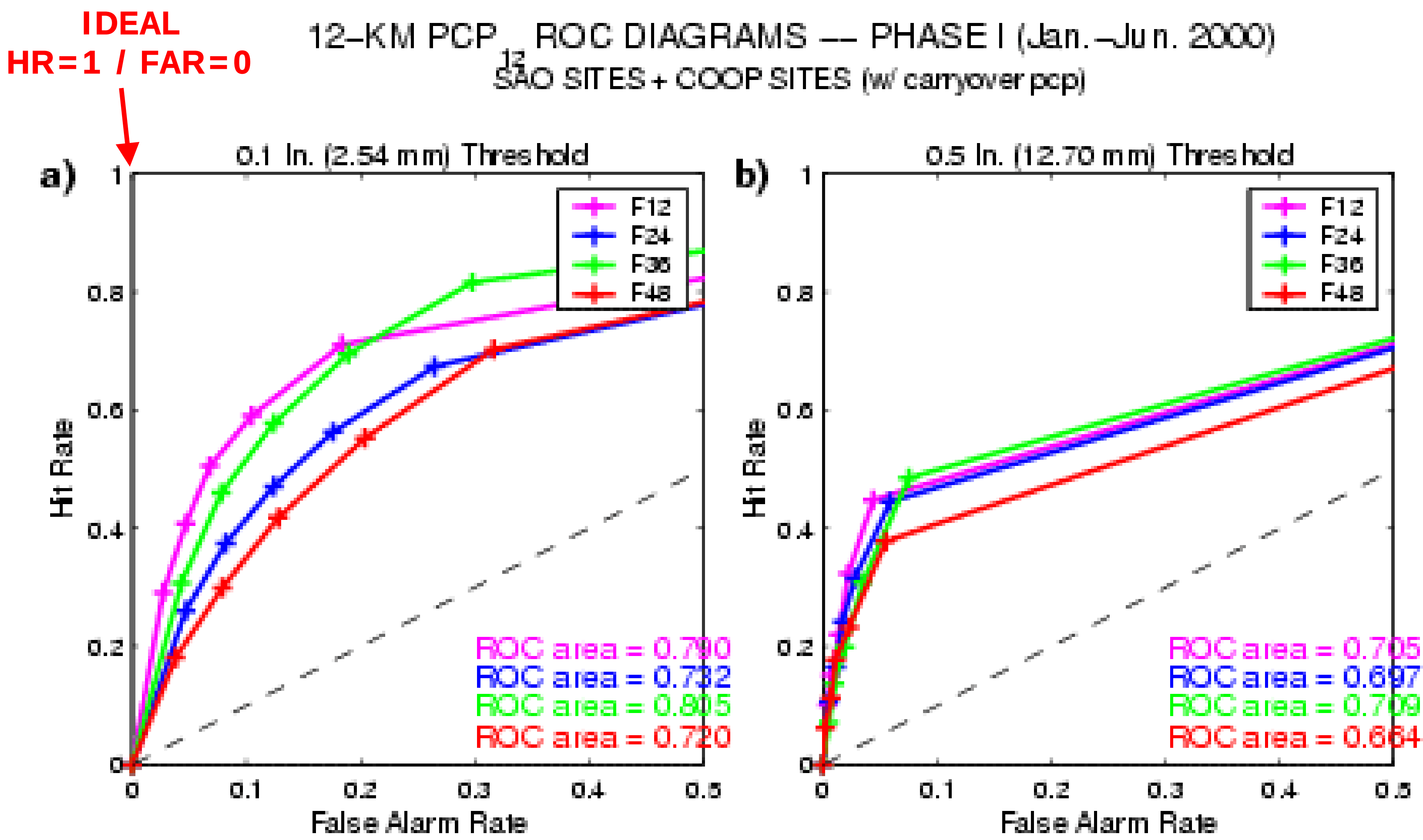
Better: ↑



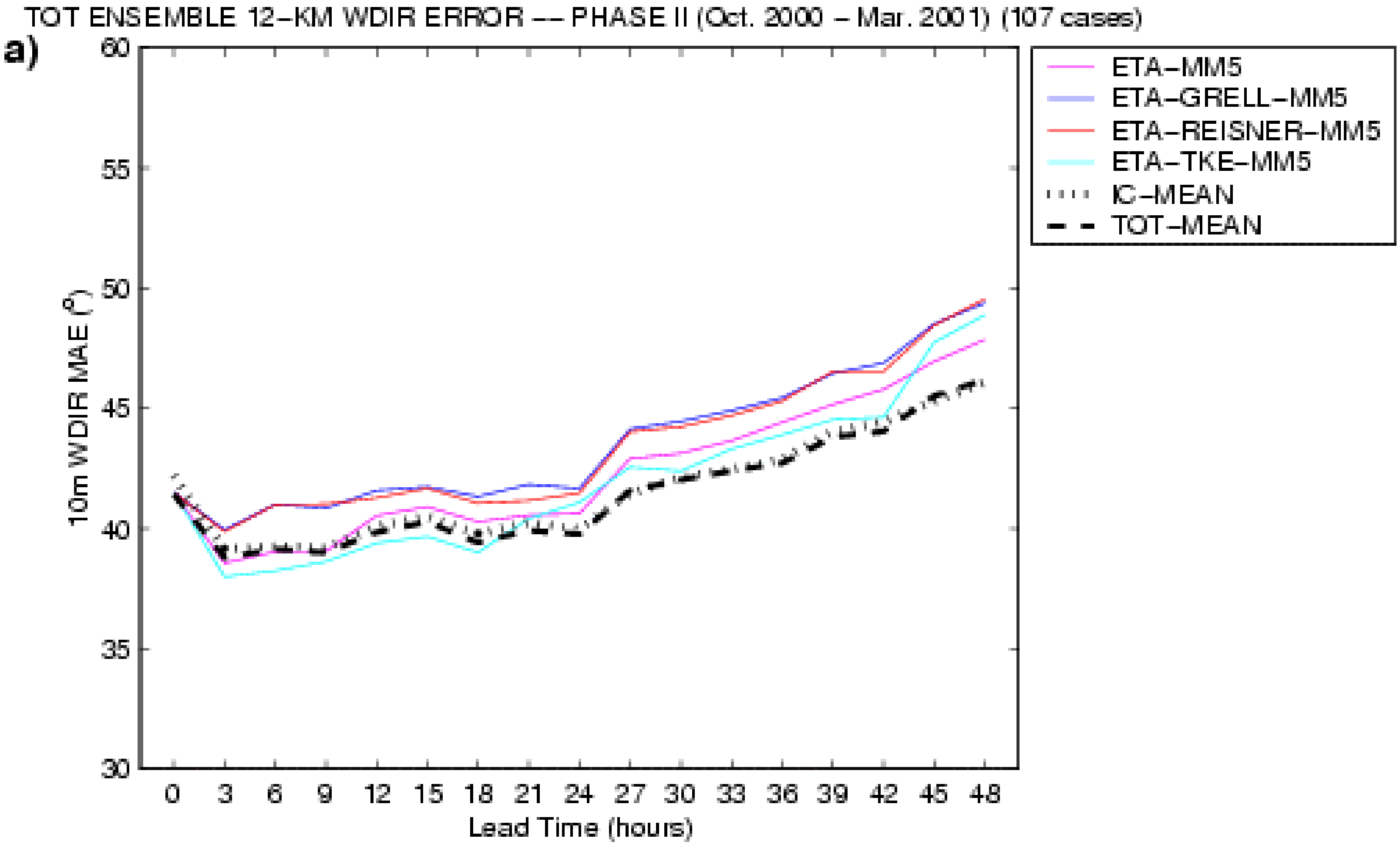
Limit of predictability for 0.1” (rain/no rain)
~ 72 h (3 days)

Phase II – less skillful

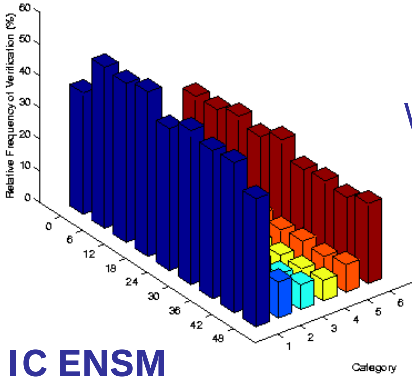
Relative Operating Characteristic (ROC) curves – measure resolution (RES)



IC MEAN vs. TOT MEAN

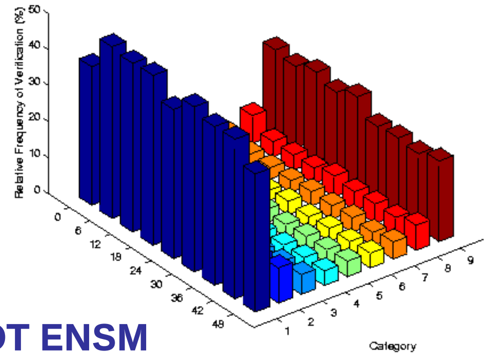


Verification Rank Histograms



WDIR

IC ENSM

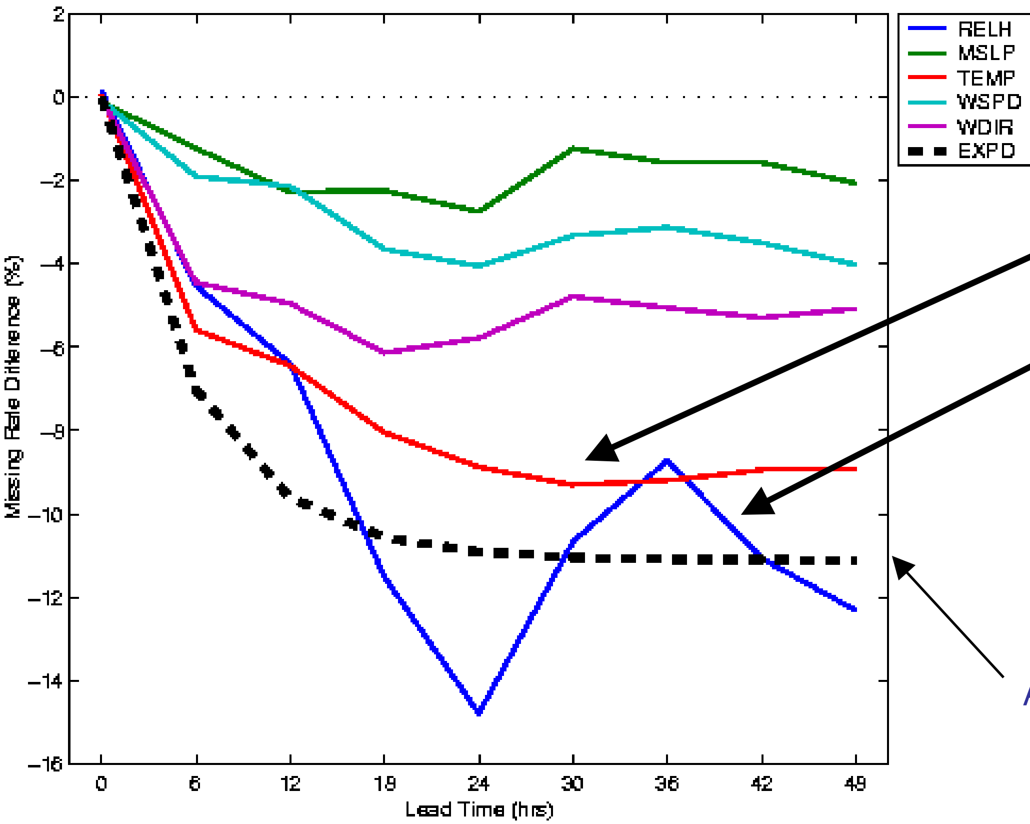


TOT ENSM

Missing rates still WAY too large!

$$>> \frac{2}{M+1}$$

Missing Rate Difference (TOT ENSM – IC ENSM)



Mixed-physics members provide most improvement for RH & TEMP.

Asymptotes to $\frac{2}{M_{TOT}+1} - \frac{2}{M_{IC}+1} = 11.1\%$

Outline

- Ensemble Forecasting Background
- Motivation for this Research
- The Approach
- Results
- **Conclusions**
- Future Work

Conclusions I

- Ensemble-mean forecasts verify better than the component forecasts over a large number of cases.
- On a case-by-case basis, ensemble-mean forecasts verify as the best forecast with about the same frequency as any member forecast.
- The 12-km ensemble-mean forecasts perform *as well as* the 4-km deterministic MM5 forecasts.
- Ensemble averaging tends to help more at higher-resolution for wind direction forecasts.
- Ensemble-mean forecasts retain many important mesoscale structures evident in the component forecasts.

Conclusions II

- The UW MM5 ensemble confirms that it can be possible to predict mesoscale forecast skill, at least for near-surface wind direction.
- Spread and error are not well correlated for cases with intermediate spread.
- Low (high) spread events are essentially more (less) predictable, since high spread/error correlations also extend to the component forecasts.
- Despite the lack of sufficient spread and the failure to adequately define the atmospheric PDF, valuable information about forecast reliability can be gleaned from the ensemble variability.

Conclusions III

- Un-calibrated PCP_{12} POP forecasts appear to have better skill than climatology until ~48h, especially for light rain events (<0.2”; <5 mm).
- Resolution (event discrimination) is good, reliability is poor.
- Limits of rain/no-rain predictability appear to be reached at 3 days.
- Predictability limits quickly drop to below 24-36h as the threshold increases above 0.2” (5 mm).
- Mixed-physics members improve the low-level temperature and moisture variability (parameters heavily influenced by sub-grid scale parameterizations)
- The effect of physics diversity on wind and sea-level pressure is much smaller

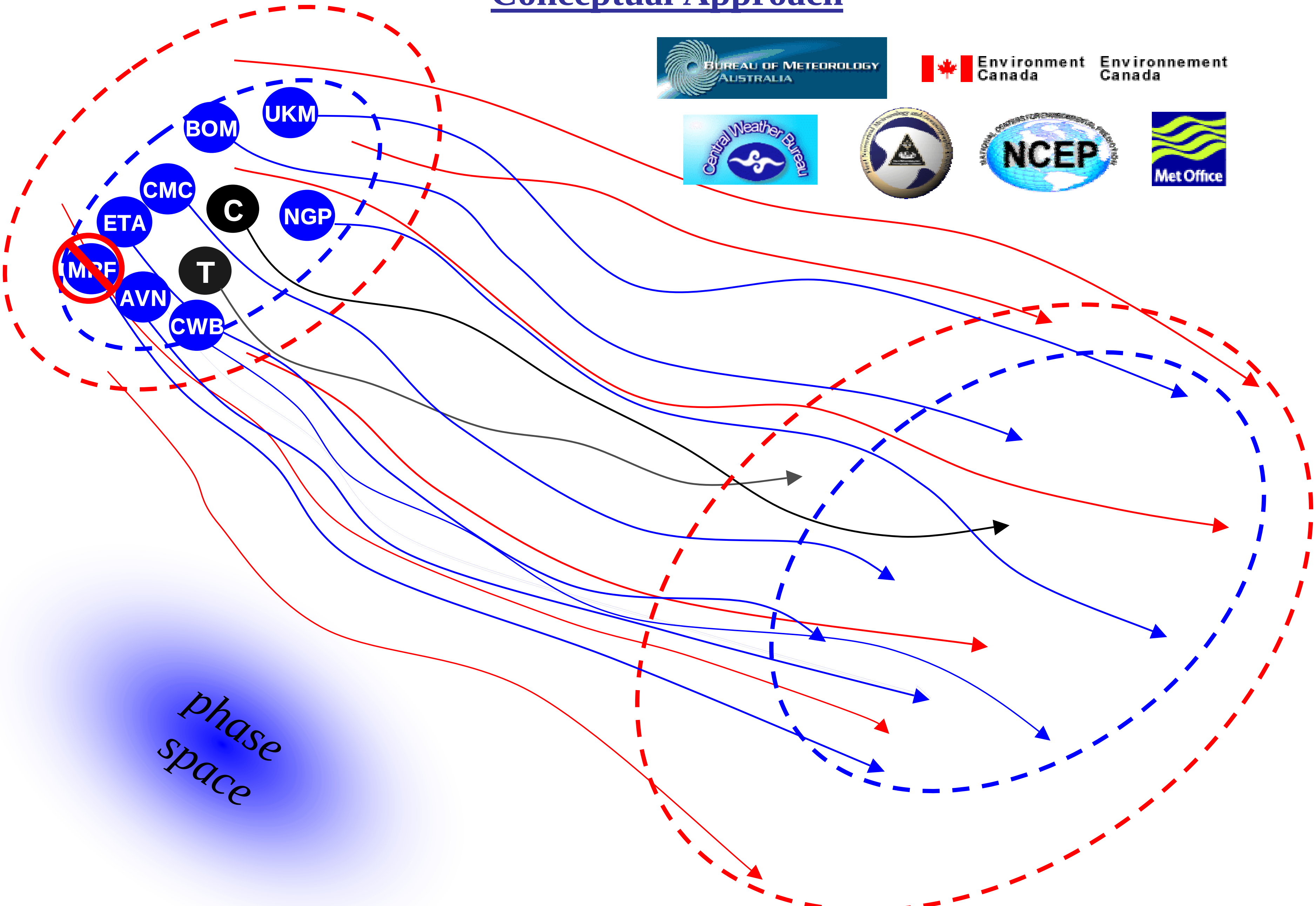
Outline

- Ensemble Forecasting Background
- Motivation for this Research
- The Approach
- Results
- Conclusions
- **Future Work**

Future Work

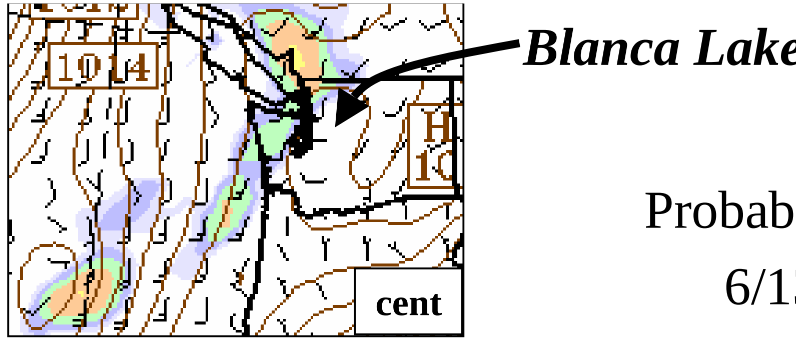
- Investigate synoptic patterns associated with extreme high & low spread.
- Investigate the spread/error relationship for other near-surface parameters, if possible.
- Use a temporal ensemble (lagged forecasts) to find any possible relationship between temporal variability and forecast reliability.
- Evaluate skill of ensemble POP forecasts compared to MOS POP
- Expand ensemble system with more operational analyses + centroid-analysis perturbations (~30-member ensemble)

Conceptual Approach



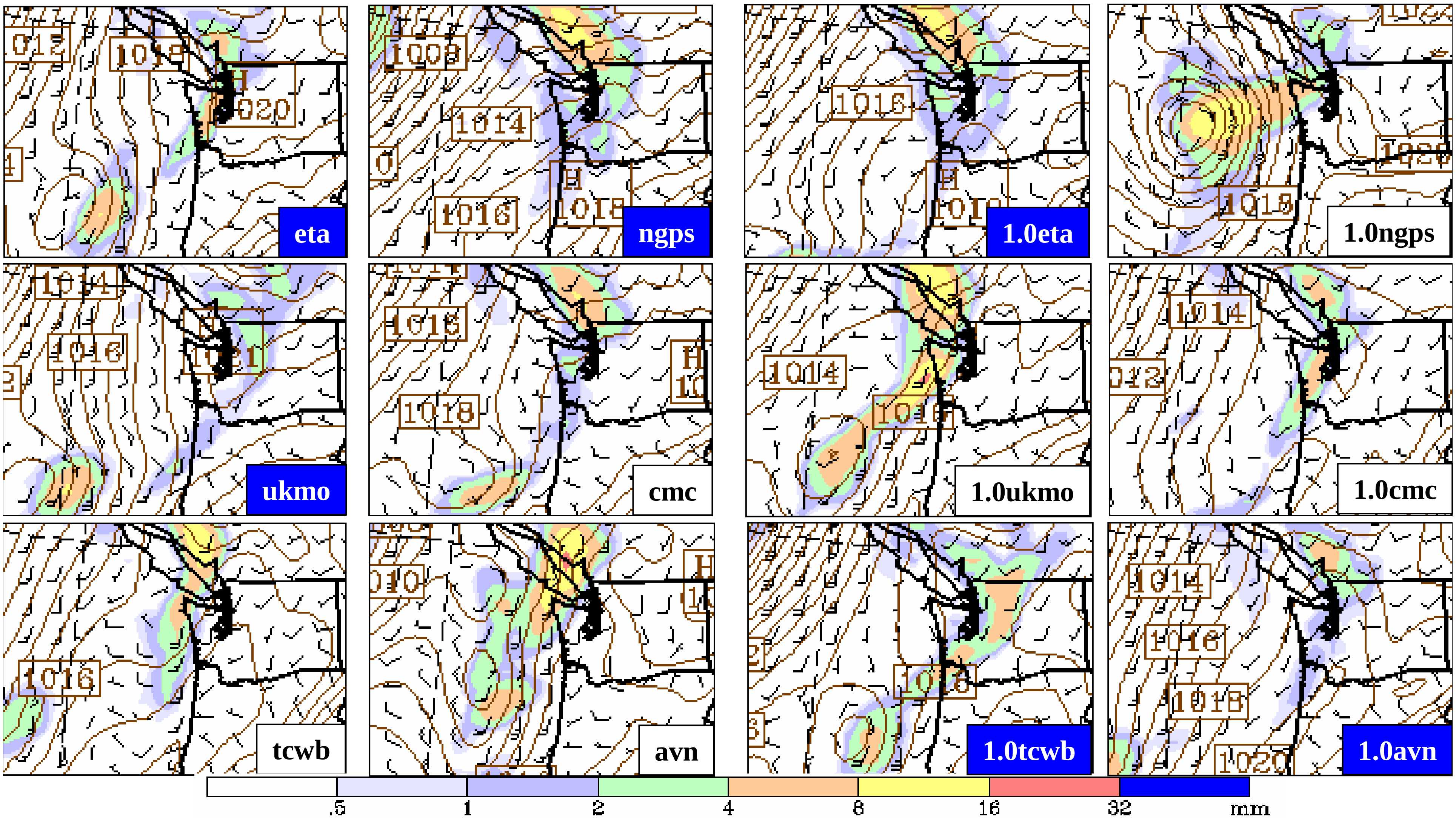
(credit to Tony Eckel)

All 13, 48h Forecasts
for slp and 6hr precip
Valid 29 Sep 00z



Probability of Precip:

6/13 = **46.2%**



Acknowledgements

- Brad Colman Sc.D. [NOAA-NWS Seattle]
- F. Anthony (Tony) Eckel [USAF & UW]
- Kari Grimit

Website

- http://www.atmos.washington.edu/~epgrimit/ensemble.cgi

Publication

- Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, in press.

[available in pdf format on website]