Diagnosis of Ensemble Forecasting Systems

M. Leutbecher

September 2009

Diagnosis of Ensemble Forecasting Systems

M. Leutbecher

September 2009

probability distribution
mean $\longleftrightarrow$ **variance**

Diagnosis of Ensemble Forecasting Systems

M. Leutbecher

ECMWF

September 2009

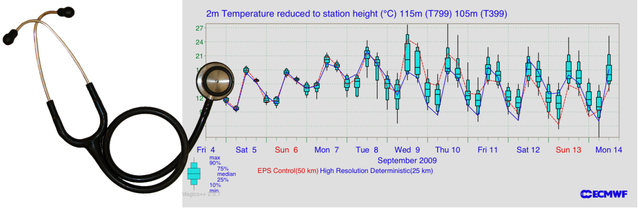
probability distribution

mean $\longleftrightarrow$ **variance**

Acknowledgements: Renate Hagedorn, Hans Hersbach, Linus Magnusson, . . . ,

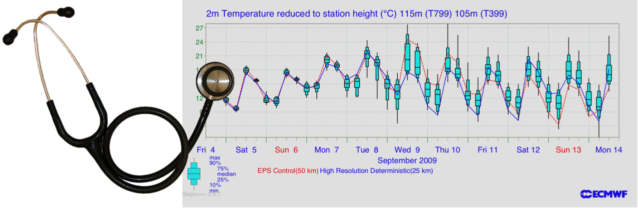
Thorpex PDP IG5: O. Talagrand, F. Atger, . . .

Diagnosis of ensemble forecast systems



Why?

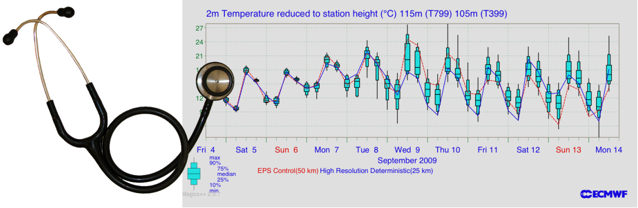
Diagnosis of ensemble forecast systems



Why?

- Aid forecast system development:
  - Quantify meteorologically relevant differences between different forecast systems.
  - Identify deficiencies
  - Provide guidance for refining the representation of initial uncertainty and model uncertainty

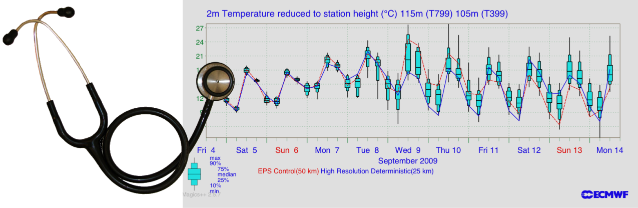
Diagnosis of ensemble forecast systems



Why?

- Aid forecast system development:
  - Quantify meteorologically relevant differences between different forecast systems.
  - Identify deficiencies
  - Provide guidance for refining the representation of initial uncertainty and model uncertainty
- Understand dynamics of (initially small) perturbations, i.e. errors, in the global circulation
  - Examine origin of large forecast errors

Diagnosis of ensemble forecast systems



Why?

- Aid forecast system development:
  - Quantify meteorologically relevant differences between different forecast systems.
  - Identify deficiencies
  - Provide guidance for refining the representation of initial uncertainty and model uncertainty
- Understand dynamics of (initially small) perturbations, i.e. errors, in the global circulation
  - Examine origin of large forecast errors

Limitations: not exhaustive, not only new developments, some of the new things are work in progress

Diagnosis of the numerical model

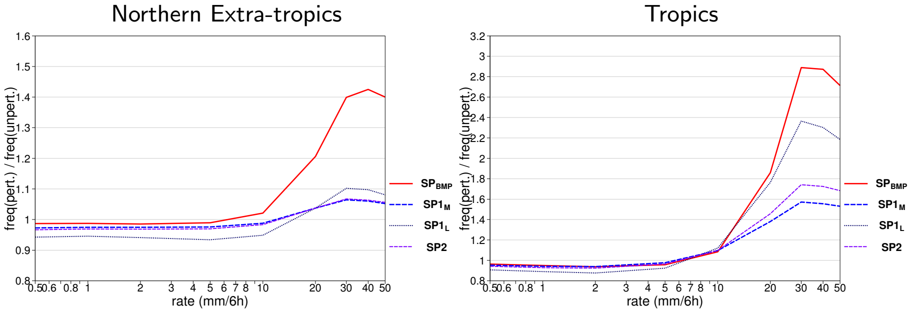
used in the ensemble forecast system

- ensemble forecast model $\neq$ model used for "deterministic" forecast: resolution, timestep, ...
- look at performance of the control forecast (unperturbed member of ensemble)
- realism of model climate of perturbed forecast model (including impact of model perturbations)

Everything as would be done for the deterministic system (except for the model perturbations).

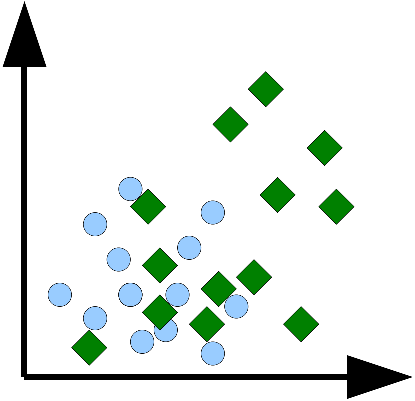
Getting the climate right: An example

- ECMWF EPS uses Stochastically Perturbed Parametrization Tendencies (SPPT) ("stochastic physics")
- Operational SPPT ($\leq 35R2$) distorts the tail of the climatological distribution of precipitation.
- A recent major revision of SPPT has improved precipitation distribution

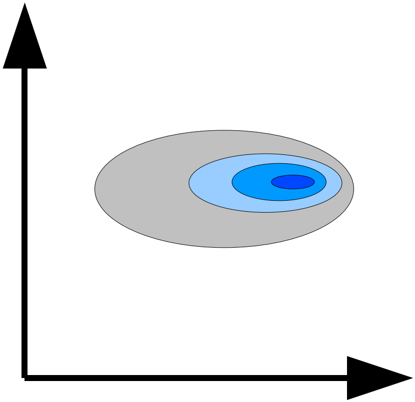


Precipitation frequency ratios between forecasts using tendency perturbations and forecasts without tendency perturbations. —— operational SPPT - - - revised SPPT

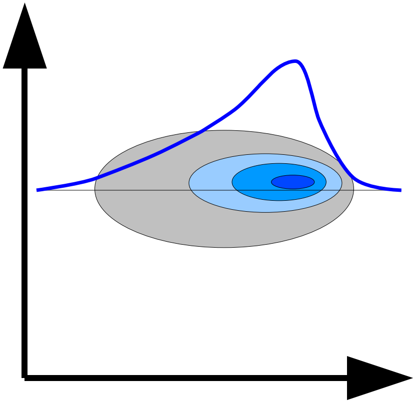
Diagnosis of deterministic forecasts



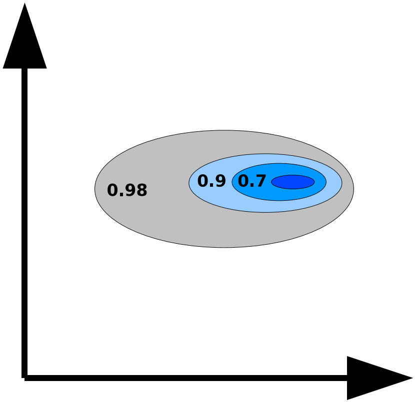
Diagnosis of probabilistic forecasts



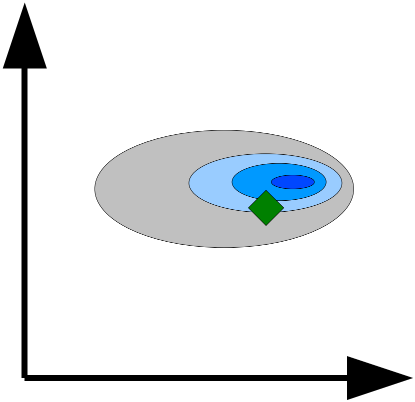
Diagnosis of probabilistic forecasts



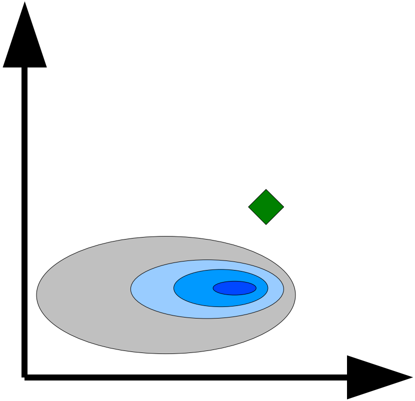
Diagnosis of probabilistic forecasts



Diagnosis of probabilistic forecasts

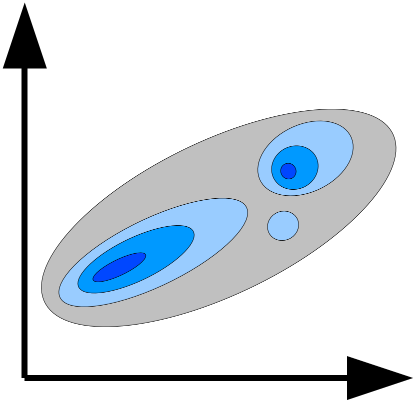


Diagnosis of probabilistic forecasts

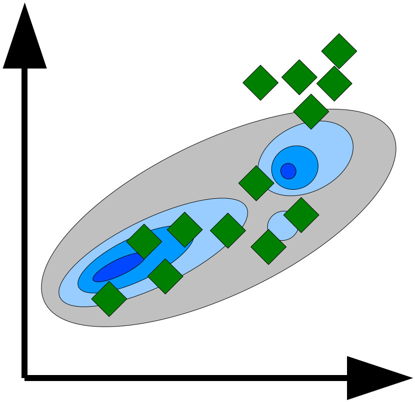


- conclusions for single cases only for exceptional failures
- forecast and verification are different objects

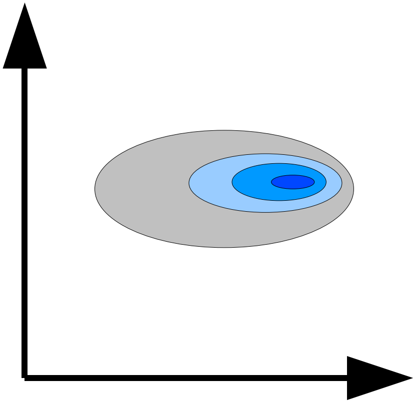
Diagnosis of probabilistic forecasts: Reliability



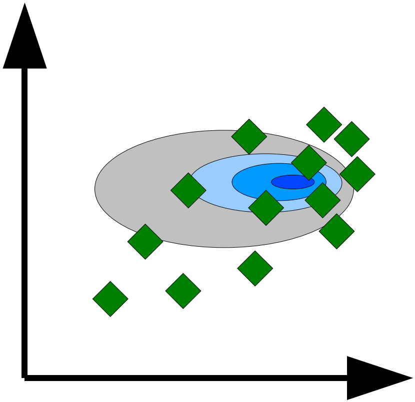
Diagnosis of probabilistic forecasts: Reliability



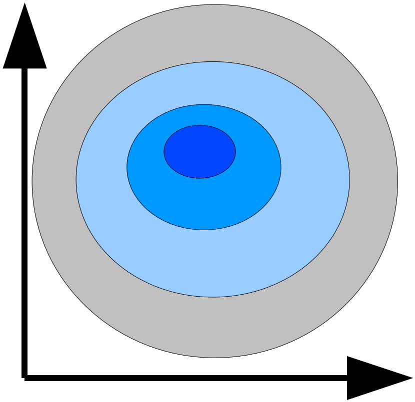
Diagnosis of probabilistic forecasts: Reliability



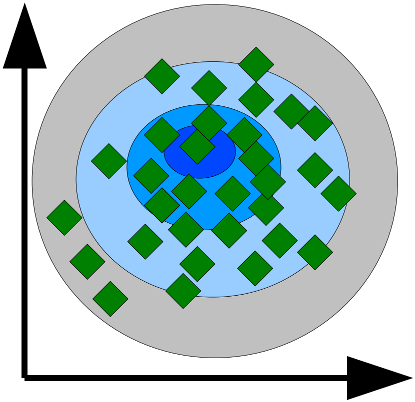
Diagnosis of probabilistic forecasts: Reliability



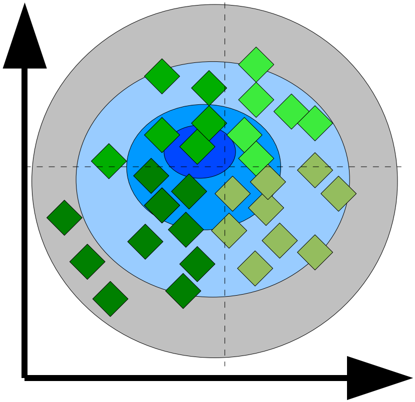
Diagnosis of probabilistic forecasts: Sharpness/Resolution



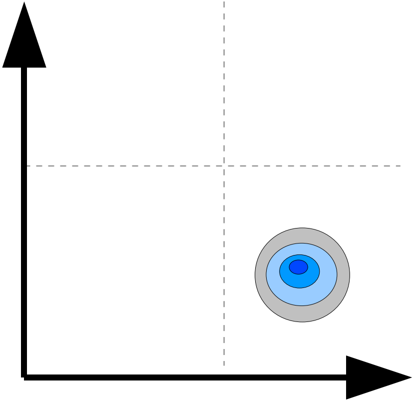
Diagnosis of probabilistic forecasts: Sharpness/Resolution



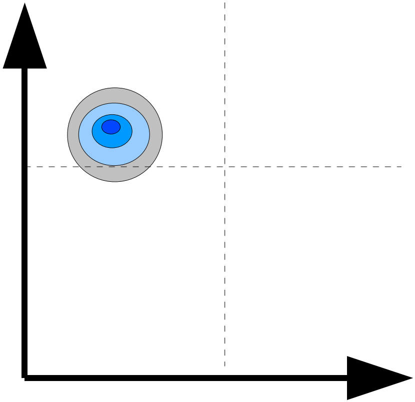
Diagnosis of probabilistic forecasts: Sharpness/Resolution



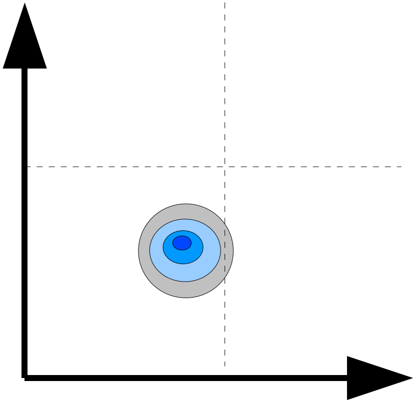
Diagnosis of probabilistic forecasts: Sharpness/Resolution



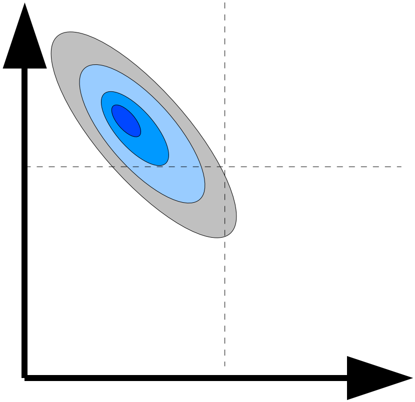
Diagnosis of probabilistic forecasts: Sharpness/Resolution



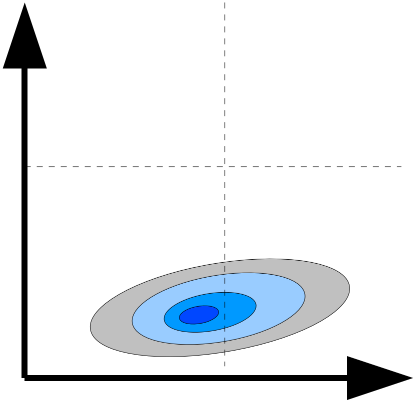
Diagnosis of probabilistic forecasts: Sharpness/Resolution



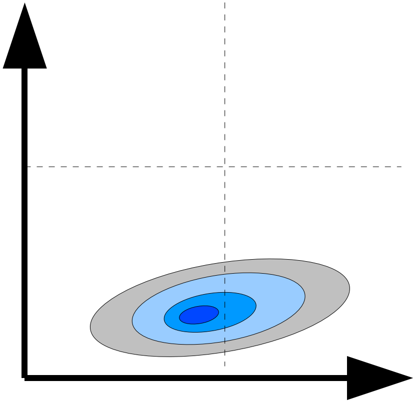
Diagnosis of probabilistic forecasts: Sharpness/Resolution



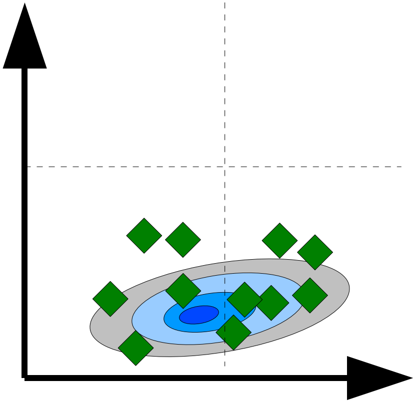
Diagnosis of probabilistic forecasts: Sharpness/Resolution



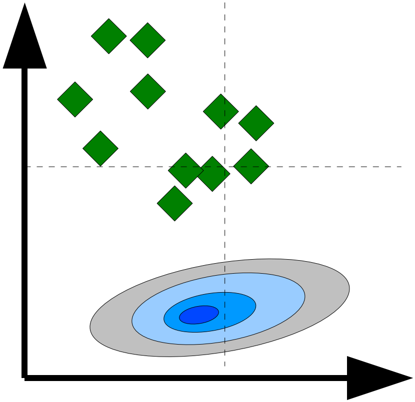
Diagnosis of probabilistic forecasts: Skill



Diagnosis of probabilistic forecasts: Skill



Diagnosis of probabilistic forecasts: Skill



Standard methods to diagnose the pdf

Impractical to assess all aspects of a multivariate probabilistic prediction

How can the assessment be simplified?

1. Limit assessment of probability distribution:
    - univariate prediction: e.g. geopotential at 500 hPa
    - binary events: does TC strike at x; prediction of a cold anomaly

Standard methods to diagnose the pdf

Impractical to assess all aspects of a multivariate probabilistic prediction

How can the assessment be simplified?

1. Limit assessment of probability distribution:
    - univariate prediction: e.g. geopotential at 500 hPa
    - binary events: does TC strike at x; prediction of a cold anomaly
2. Use summary measures of the overall quality of a predicted pdf (or some aspect of it):
    - Skill of the ensemble mean
    - Match between Ens. Mean RMS error and Ensemble Stdev. (reliability)

Standard methods to diagnose the pdf

Impractical to assess all aspects of a multivariate probabilistic prediction

How can the assessment be simplified?

1. Limit assessment of probability distribution:
   - univariate prediction: e.g. geopotential at 500 hPa
   - binary events: does TC strike at x; prediction of a cold anomaly
2. Use summary measures of the overall quality of a predicted pdf (or some aspect of it):
   - Skill of the ensemble mean
   - Match between Ens. Mean RMS error and Ensemble Stdev. (reliability)
   - Rank Histogram: reliability
   - Brier score, (Continuous) Ranked Probability Score: reliability and resolution

Standard methods to diagnose the pdf

Impractical to assess all aspects of a multivariate probabilistic prediction

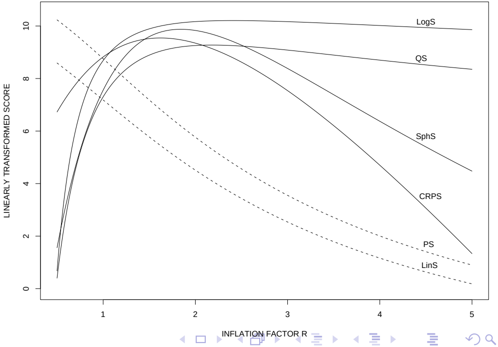
How can the assessment be simplified?

1. Limit assessment of probability distribution:
   - univariate prediction: e.g. geopotential at 500 hPa
   - binary events: does TC strike at x; prediction of a cold anomaly
2. Use summary measures of the overall quality of a predicted pdf (or some aspect of it):
   - Skill of the ensemble mean
   - Match between Ens. Mean RMS error and Ensemble Stdev. (reliability)
   - Rank Histogram: reliability
   - Brier score, (Continuous) Ranked Probability Score: reliability and resolution (decomposition!)
   - Relative Operating Characterisitic (ROC): (discrimination)
   - Logarithmic Score (Ignorance): reliability and resolution

Proper scores

- **strictly proper** implies that optimizing the score leads to the correct probability distribution
- optimization of a score that is not proper is likely to lead to a wrong distribution
- concise mathematical definitions of *proper* and *strictly proper* are available (see Gneiting and Raftery, 2004)
- examples of proper scores: BS, RPS, CRPS, logarithmic score

Fig. 1,
Gneiting and Raftery (2004)



Spread-error relationship and ensemble size

Assume a perfectly reliable (statistically consistent) M-member ensemble: Ens. members $x_j, j = 1, \ldots, M$ and truth y are independent draws from a distribution with mean μ and variance σ^2.

Spread-error relationship and ensemble size

Assume a perfectly reliable (statistically consistent) M-member ensemble: Ens. members $x_j, j = 1, \ldots, M$ and truth y are independent draws from a distribution with mean μ and variance σ^2.

Expected squared error of ensemble mean

$$\mathbb{E}\left(\underbrace{\frac{1}{M}\sum_{j=1}^{M} x_j}_{\mu+\text{sampling error}} - y\right)^2 = \left(1 + \frac{1}{M}\right)\sigma^2$$

Spread-error relationship and ensemble size

Assume a perfectly reliable (statistically consistent) M-member ensemble: Ens. members $x_j, j = 1, \ldots, M$ and truth y are independent draws from a distribution with mean μ and variance σ^2.

Expected squared error of ensemble mean

$$\mathbb{E}\left(\underbrace{\frac{1}{M}\sum_{j=1}^{M} x_j}_{\mu+\text{sampling error}} - y\right)^2 = \left(1+\frac{1}{M}\right)\sigma^2$$

Expected ensemble variance

$$\mathbb{E}\,\frac{1}{M}\sum_{j=1}^{M}\left(x_j - \underbrace{\frac{1}{M}\sum_{k=1}^{M} x_k}_{\neq\mu,\ \text{contains}\ x_j}\right)^2 = \left(1-\frac{1}{M}\right)\sigma^2$$

Spread-error relationship and ensemble size

Perfectly reliable M-member ensemble:
Ens. members $x_j, j = 1, \ldots, M$ and truth y are independent draws from a distribution with mean μ and variance σ^2.

For large ensembles, e.g. $M = 50$,

$$\overline{\text{ensemble variance}} = \overline{\text{squared ensemble mean error}}$$

in practice.

Spread-error relationship and ensemble size

Perfectly reliable M-member ensemble:
Ens. members $x_j, j = 1, \ldots, M$ and truth y are independent draws from a distribution with mean μ and variance σ^2.

For large ensembles, e.g. $M = 50$,

$$\overline{\text{ensemble variance}} = \overline{\text{squared ensemble mean error}}$$

in practice.

For smaller ensemble size, e.g. $M \leq 20$,

$$\left(1 - \frac{1}{M}\right)^{-1} \overline{\text{ens. variance}} = \left(1 + \frac{1}{M}\right)^{-1} \overline{\text{squared ens. mean error}}$$

Spread versus error in 2002

500 hPa geopotential height, N.-Hem. extra-tropics

dashed: ens. stdev.
solid: EM RMSE
10 member ensembles

ECMWF(T_L255L40), MSC, NCEP

May–July 2002 *Fig. 5, Buizza et al. 2005*

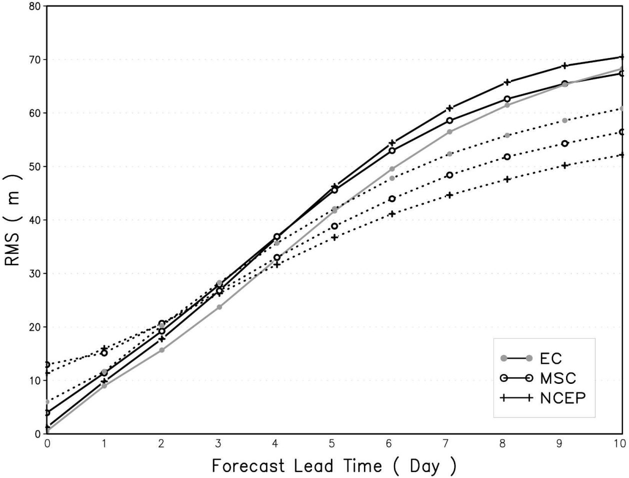


FIG. 5. May–Jun–Jul 2002 average rms error of the ensemble mean (solid lines) and ensemble standard deviation (dotted lines) of the EC-EPS (gray lines with full circles), the MSC-EPS (black lines with open circles), and the NCEP-EPS (black lines with crosses). Values refer to the 500-hPa geopotential height over the Northern Hemisphere latitudinal band 20°–80°N.

Spread versus error in 2007

500 hPa geopotential, N.-Hem. extra-tropics

ECMWF, cycle 32r2 cycle 32r3

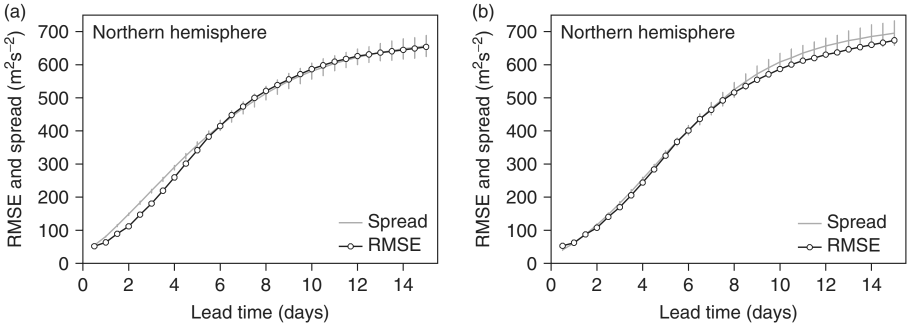


Fig. 12, Bechtold et al. 2008

50 member, T_L399L62

69 cases, June-Sept. 2007

Spread versus error in 2007

500 hPa geopotential, N.-Hem. extra-tropics

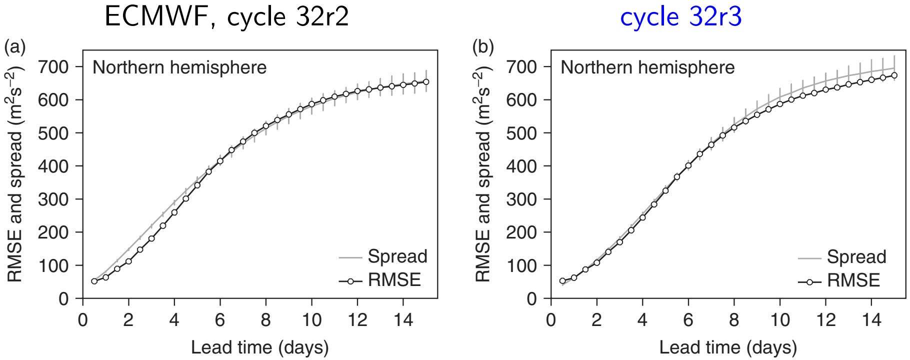


Fig. 12, Bechtold et al. 2008

50 member, T_L399L62

69 cases, June-Sept. 2007

Improved match due to revised model physics together with a 30% reduction of the initial perturbation amplitude

Spread versus error in 2007

500 hPa geopotential, N.-Hem. extra-tropics

ECMWF, cycle 32r2 cycle 32r3

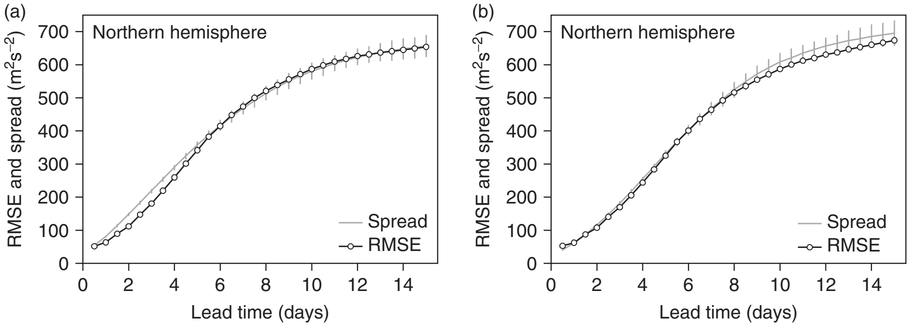


Fig. 12, Bechtold et al. 2008

50 member, $T_L399L62$

69 cases, June-Sept. 2007

Improved match due to revised model physics together with a 30% reduction of the initial perturbation amplitude

Time to retire?

Explore more directions in phase space

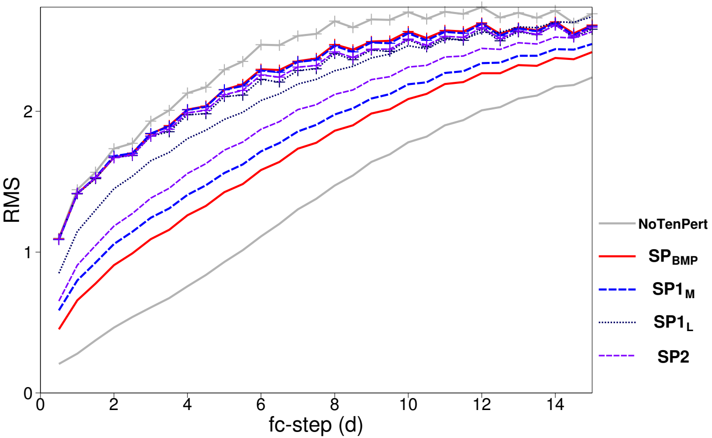
- Agreement for other variables (e.g. T850)?
- Other regions (e.g. tropics)?
- Detailed geographical distribution of spread?
- Different spatial scales?

Example: v850 tropics (20°S–20°N)

SPPT revision

Ens. Mean RMSE —+—

Ens. Stdev ——

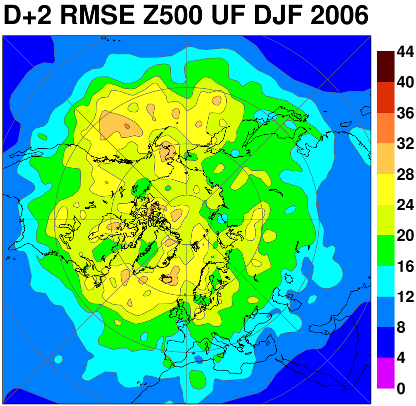


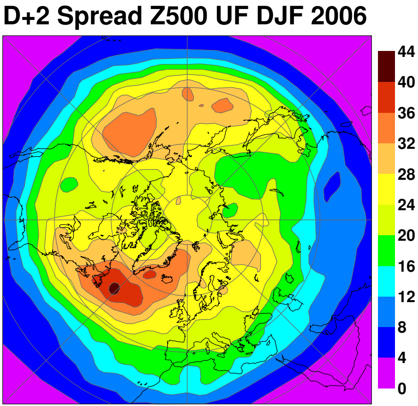
—— operational SPPT - - - revised SPPT

40 cases Nov/Dec 2007 + Jul/Aug 2008, T_L399(255 from D+10)

Scale-dependent spread-error diagnostic

spread map D+2: unfiltered fields



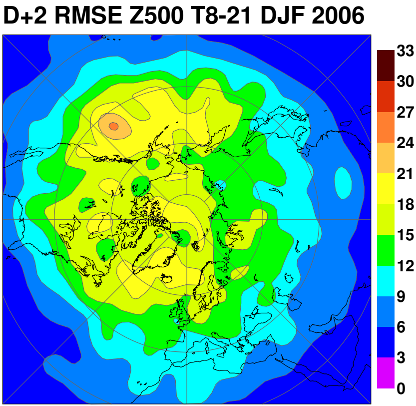


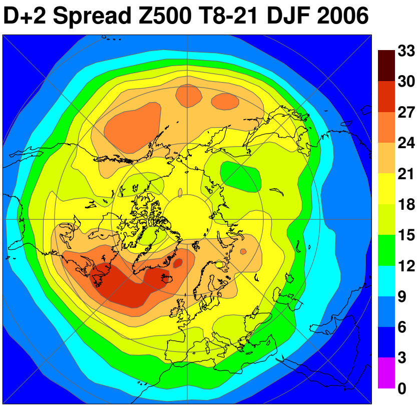
region	spread : RMSE
20°–90°N	1.16
35°-65°N	1.22

Jung and Leutbecher (2008)

Scale-dependent spread-error diagnostic

spread map D+2: total wavenumber 8–21





region	spread : RMSE
20°–90°N	1.32
35°-65°N	1.37

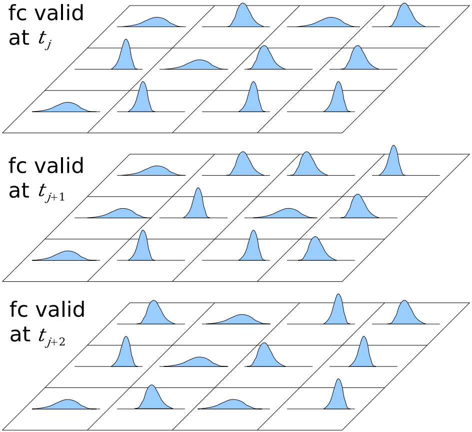
Jung and Leutbecher (2008)

Assessing flow-dependent and data-dependent variations in pdf-shape

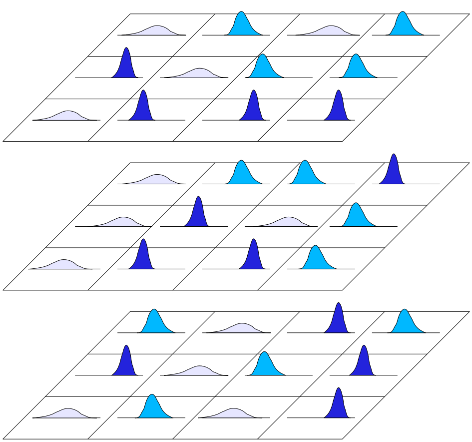
sample size limited: therefore initially focus on 2nd moment of pdf, i.e. variance

1. Stratification by ensemble standard deviation: spread-reliability
2. Modified event definition: EM error $> \theta$, where threshold θ depends on a "climatological" stdev of the EM error
3. Gaussian centred on CF or EM as reference. Stdev of Gaussian can vary geographically and seasonally.

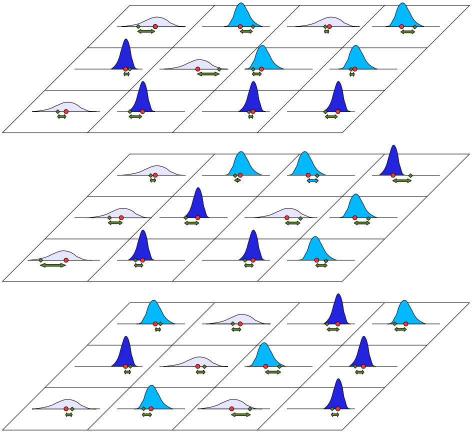
spread-reliability: methodology



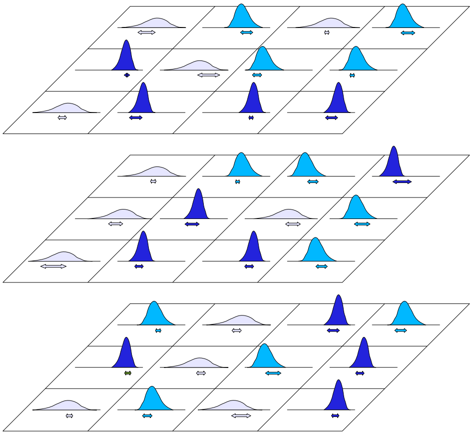
spread-reliability: methodology



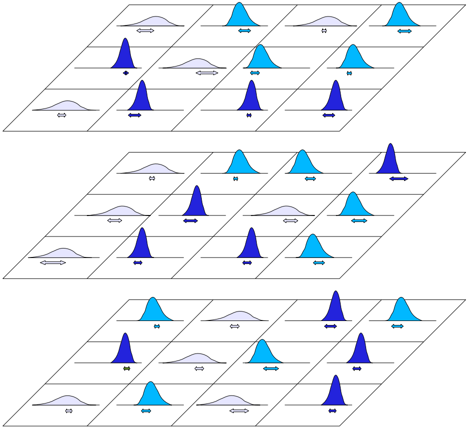
spread-reliability: methodology

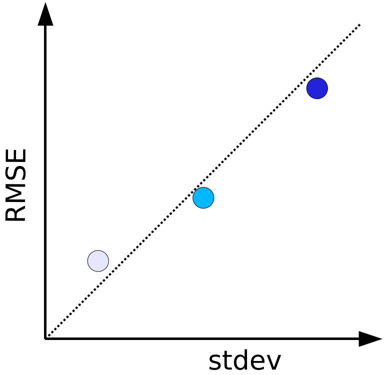


spread-reliability: methodology

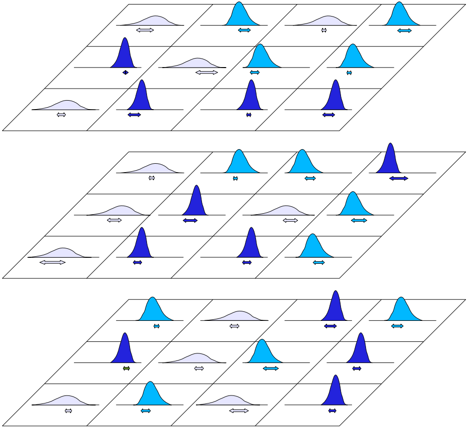


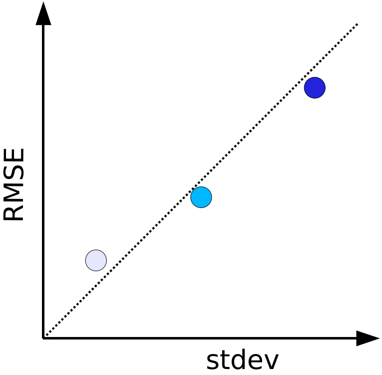
spread-reliability: methodology





spread-reliability: methodology





see also Leutbecher and Palmer (2008); Leutbecher et al. (2007)

spread-reliability: Z500 DJF06/07

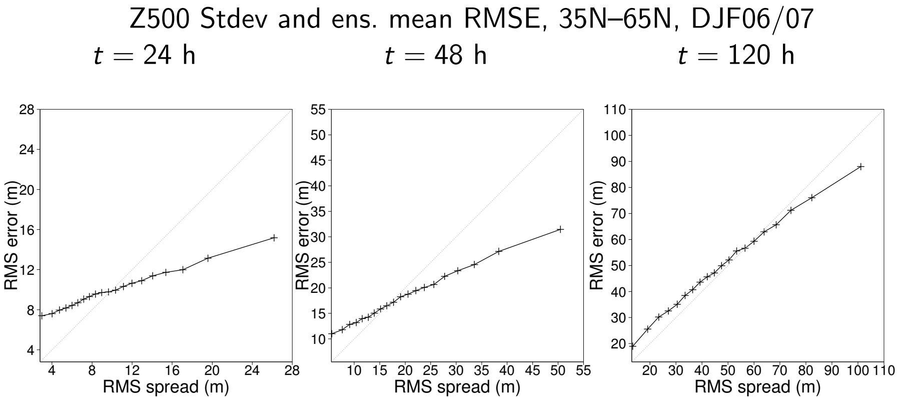


Fig. 7 from Leutbecher, Buizza & Isaksen (2007)

spread-reliability: Z500 DJF06/07

Z500 Stdev and ens. mean RMSE, 35N–65N, DJF06/07

$t = 24$ h　　$t = 48$ h　　$t = 120$ h

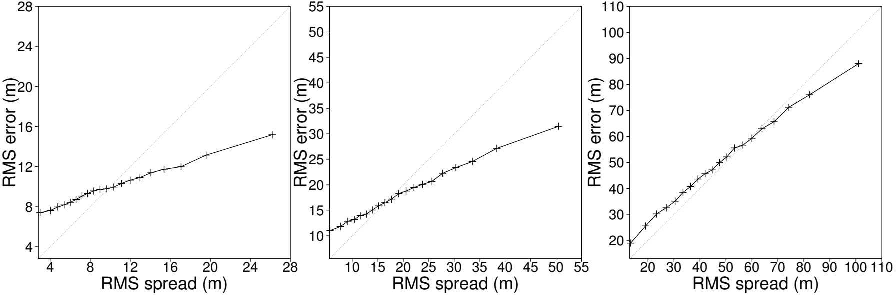


Fig. 7 from Leutbecher, Buizza & Isaksen (2007)

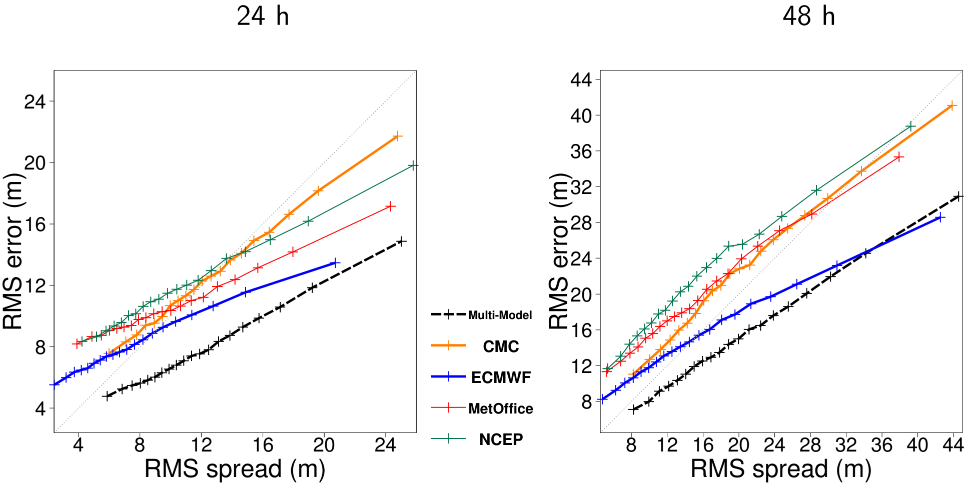
deficiency at the early forecast ranges; reliability improves with lead time

Comparison with other ensembles in TIGGE

- data provided by Renate Hagedorn
- direct model output
- verified with quasi-independent analysis: ERA-Interim
- period: DJF2008/2009 (0 UTC, 90 start dates)
- region: N.-Hem midlatitudes (35°–65°N)

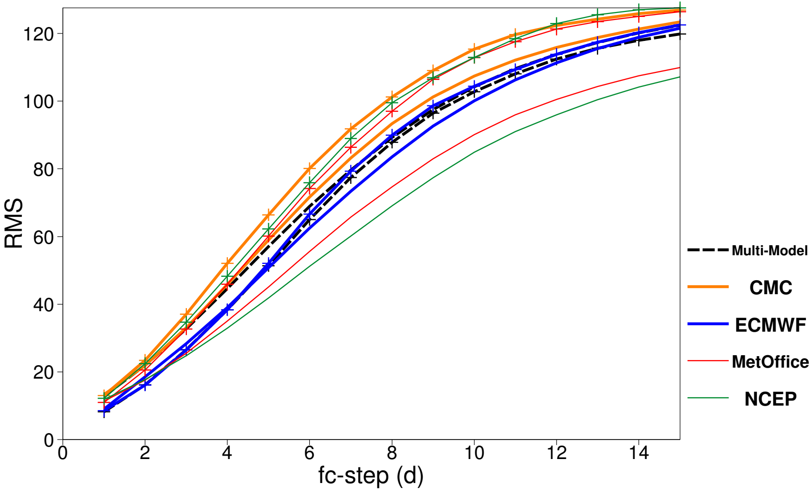
Spread-reliability: GH 500 hPa

TIGGE comparison



Ens. stdev and EM RMS error: 500 hPa geopotential

TIGGE comparison

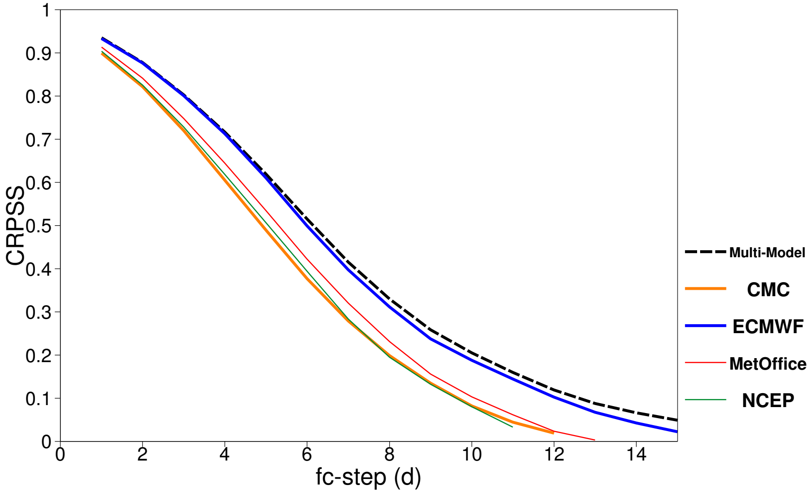


Ens. Mean RMSE

Ens. stdev

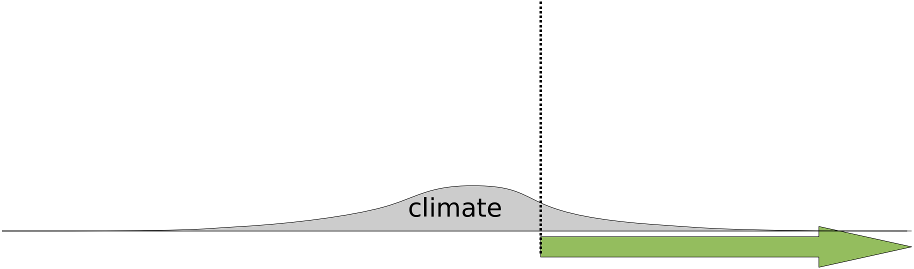
CRPS: 500 hPa geopotential

TIGGE comparison



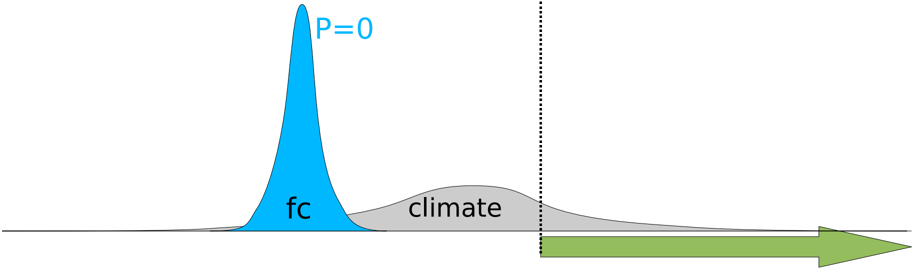
Binary events based on the climate

consider a short lead time . . .



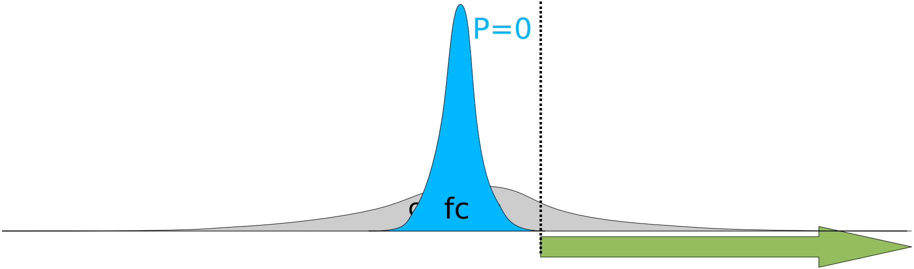
Binary events based on the climate

consider a short lead time . . .



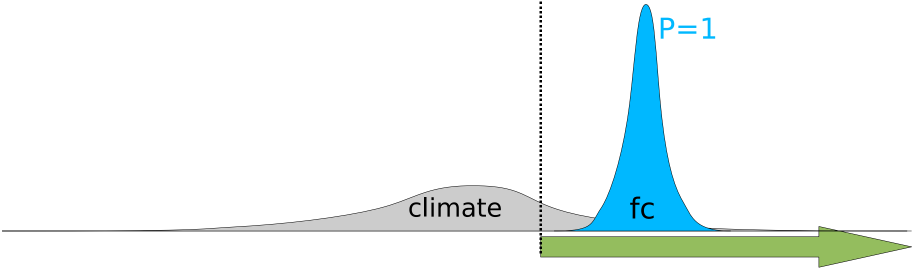
Binary events based on the climate

consider a short lead time . . .



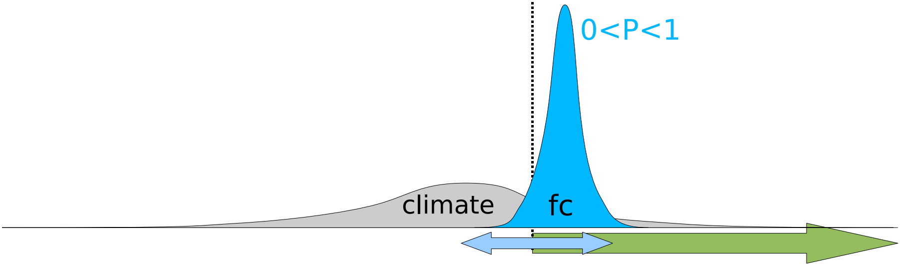
Binary events based on the climate

consider a short lead time . . .



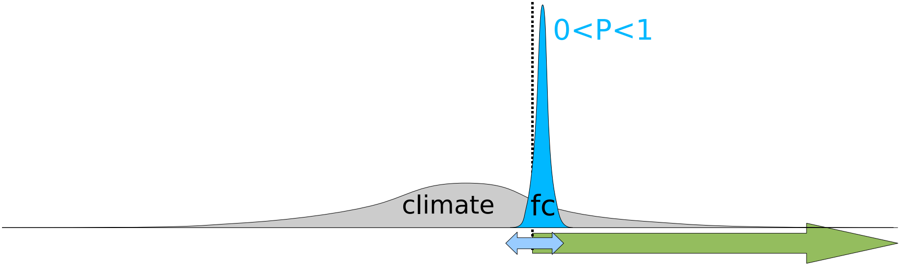
Binary events based on the climate

consider a short lead time . . .

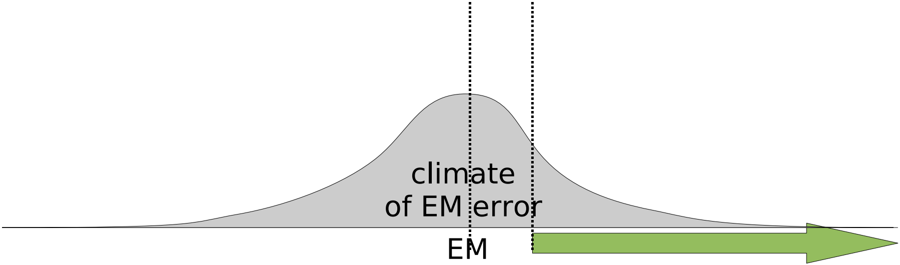


Binary events based on the climate

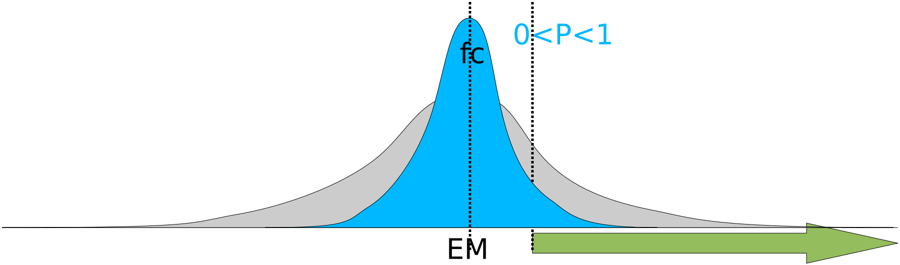
consider an *even shorter* lead time . . .



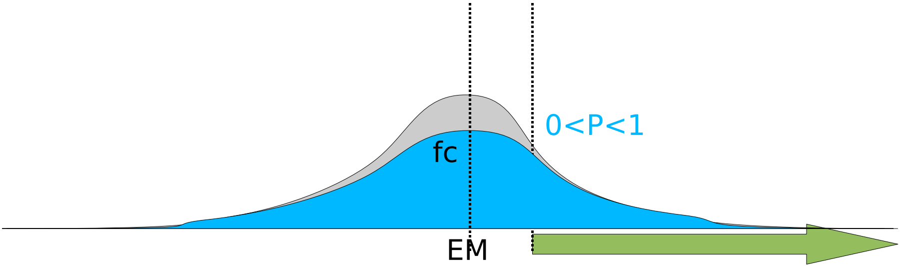
Binary events based on Ens. mean and its error climate



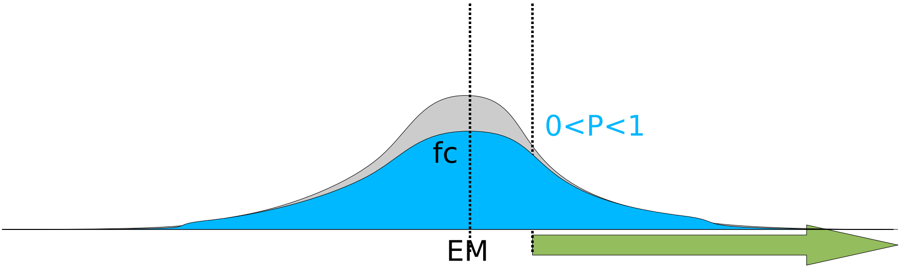
Binary events based on Ens. mean and its error climate



Binary events based on Ens. mean and its error climate

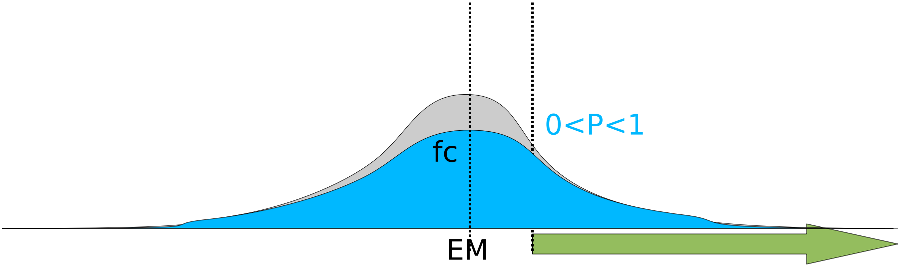


Binary events based on Ens. mean and its error climate



scales naturally with lead time, expect to be better suited to diagnose skill of variations in pdf-shape

Binary events based on Ens. mean and its error climate



scales naturally with lead time, expect to be better suited to diagnose skill of variations in pdf-shape

can also use CF and climate of CF errors . . .

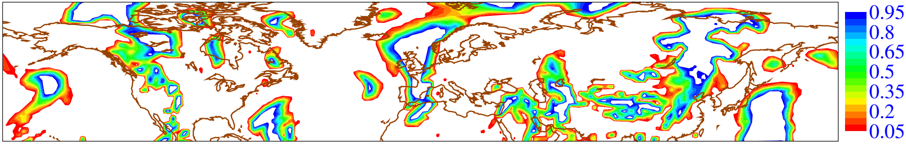
An error climatology based on reanalyses and reforecasts

- reforecast started from ERA-40 and operational analyses (reforecasts from ERA-Interim operational since March 2009)
- 5 members (CF + 4 PF) $\Rightarrow$ Ens. mean slightly less accurate
- 9 weeks centred on week of interest
- 18 years, once weekly $\Rightarrow 18 \times 9 = 162$ errors
- errors for climatology computed with ERA-Interim analyses
- verification for DJF09 with operational analyses

Probability of different kinds of events

48-hour fc of 850 hPa meridional velocity
valid at 0 UTC on 31 January 2009

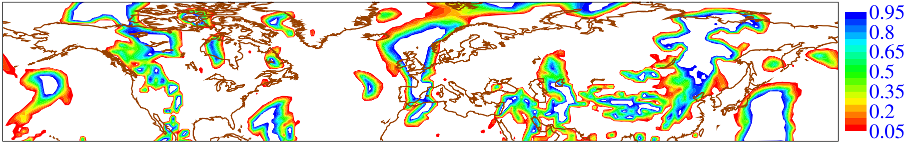
$P(x > \mu_{\text{clim}} + \sigma_{\text{clim}})$



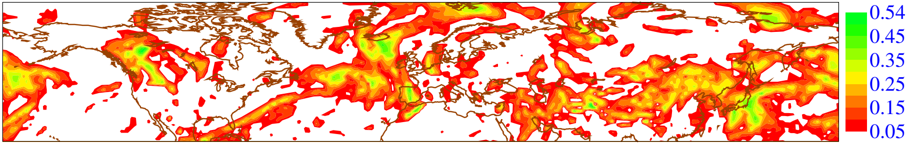
Probability of different kinds of events

48-hour fc of 850 hPa meridional velocity
valid at 0 UTC on 31 January 2009

$P(x > \mu_{\text{clim}} + \sigma_{\text{clim}})$



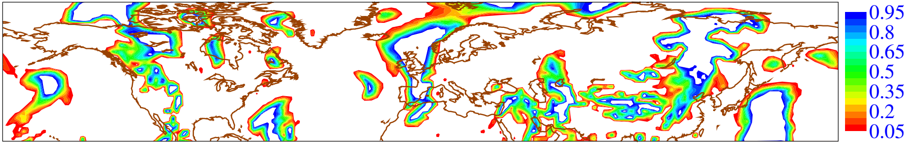
$P(x > \text{EM} + \sigma_{\text{err}})$ expect on average 0.16



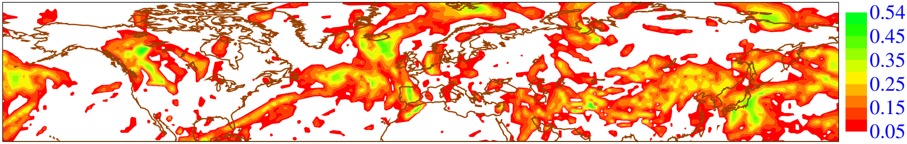
Probability of different kinds of events

48-hour fc of 850 hPa meridional velocity
valid at 0 UTC on 31 January 2009

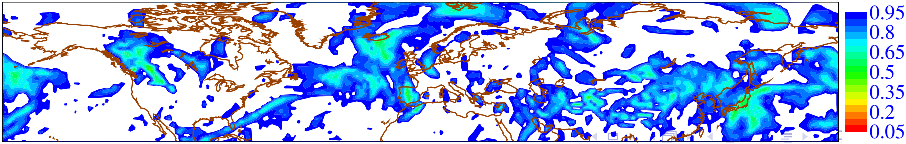
$P(x > \mu_{\text{clim}} + \sigma_{\text{clim}})$



$P(x > \text{EM} + \sigma_{\text{err}})$ expect on average 0.16



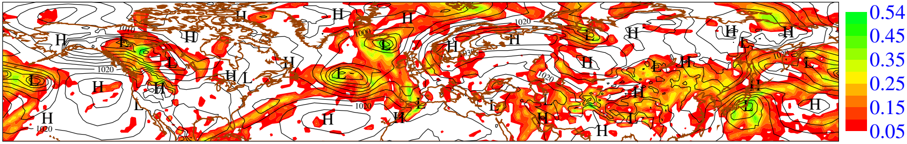
$P(x > \text{EM} - \sigma_{\text{err}})$ expect on average 0.84



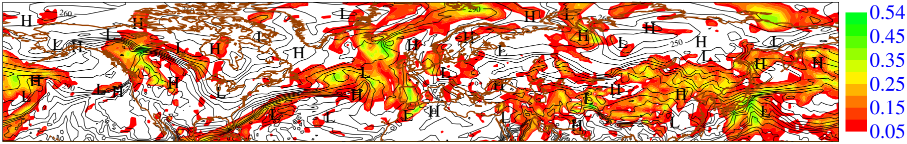
Probability of different kinds of events (2)

48-hour fc of 850 hPa meridional velocity
valid at 0 UTC on 31 January 2009

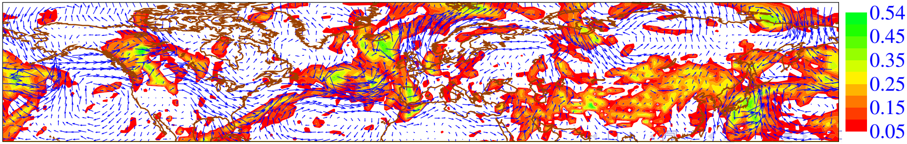
$P(x > \mathrm{EM} + \sigma_{\mathrm{err}})$ and mslp



$P(x > \mathrm{EM} + \sigma_{\mathrm{err}})$ and θ_e at 925 hPa



$P(x > \mathrm{EM} + \sigma_{\mathrm{err}})$ and wind at 850 hPa



Probabilistic scores for new types of events

- **Brier Score**: $\overline{(p-o)^2}$

Probabilistic scores for new types of events

- **Brier Score**: $\overline{(p-o)^2}$
- **Logarithmic Score** (Ignorance):
  $-\overline{(o\log(p^{(W)}) + (1-o)\log(1-p^{(W)}))}$, where

$$p^{(W)}(n) = \frac{n+2/3}{M+4/3} \in \left[\frac{2}{3M+4}, \frac{3M+2}{3M+4}\right]$$

with n being the number of members predicting the event and M being the ensemble size. The $p^{(W)}(n)$ are known as Tukey plotting position; cf. also Cromwell's rule and Wilks (2006).

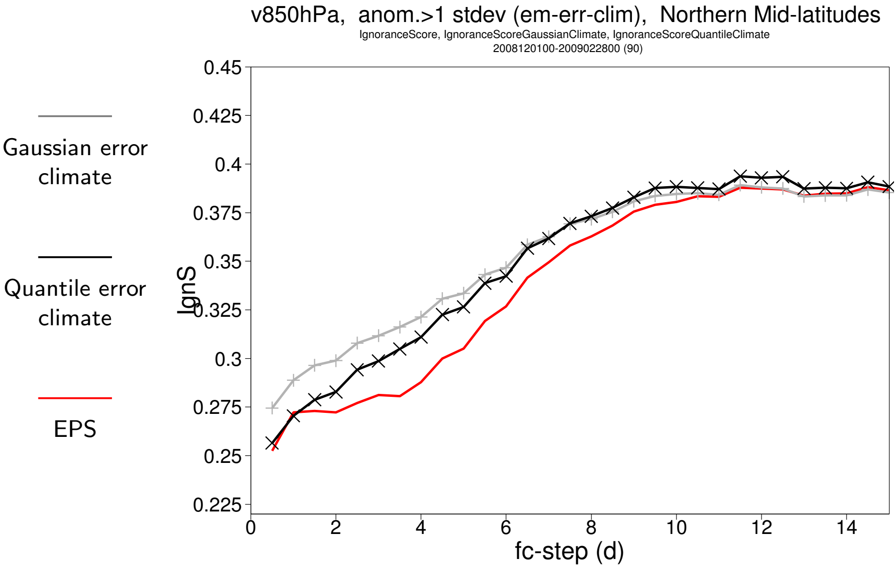
Probabilistic scores for new types of events

- **Brier Score**: $\overline{(p-o)^2}$
- **Logarithmic Score** (Ignorance):
$\overline{-(o\log(p^{(W)}) + (1-o)\log(1-p^{(W)}))}$, where

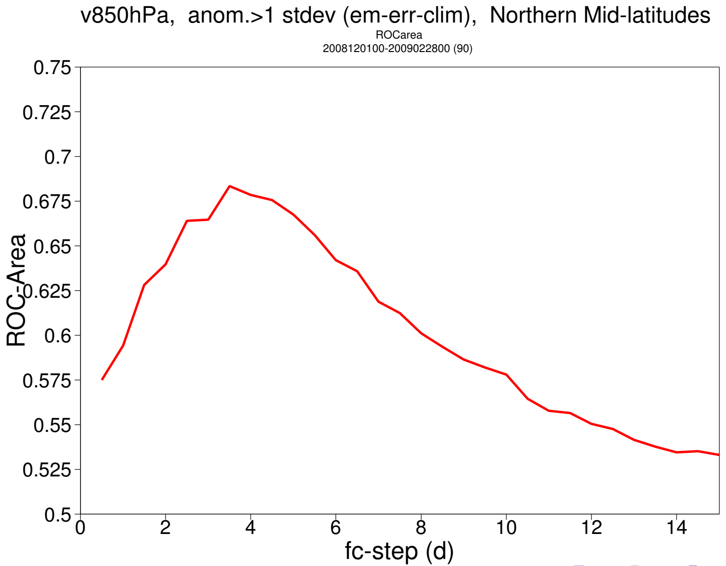
$$p^{(W)}(n) = \frac{n+2/3}{M+4/3} \in \left[\frac{2}{3M+4}, \frac{3M+2}{3M+4}\right]$$

with n being the number of members predicting the event and M being the ensemble size. The $p^{(W)}(n)$ are known as Tukey plotting position; cf. also Cromwell's rule and Wilks (2006).
- **ROC-area**: $\int_0^1 H \,\mathrm{d}F \in [0.5, 1]$, where H and F denote Hit Rate and False Alarm Rate, respectively.

Logarithmic Score for $x > \mathrm{EM} + \sigma_{\mathrm{err}}$



Area under the ROC for $x > \mathrm{EM} + \sigma_{\mathrm{err}}$



Intermediate Summary: Events relative to EM/CF

- Overall, scores (BS, IgnS, ROC-area) indicate that EPS has more skill in predicting variations in pdf shape than climatological error pdf
  - However, additional EPS skill tends to be relatively small initially.
  - It increases to max typically around $t \approx 4 \pm 2\,\mathrm{d}$.
  - Then, additional skill gradually decreases
- Similar results for T850, Z500, and also for MAM09, and for verification with ERA-Interim analyses (not shown)

Intermediate Summary: Events relative to EM/CF

- Overall, scores (BS, IgnS, ROC-area) indicate that EPS has more skill in predicting variations in pdf shape than climatological error pdf
  - However, additional EPS skill tends to be relatively small initially.
  - It increases to max typically around $t \approx 4 \pm 2\,\mathrm{d}$.
  - Then, additional skill gradually decreases
- Similar results for T850, Z500, and also for MAM09, and for verification with ERA-Interim analyses (not shown)
- Initial skill increase consistent with fact that spread-error reliability improves with lead time

Intermediate Summary: Events relative to EM/CF

- Overall, scores (BS, IgnS, ROC-area) indicate that EPS has more skill in predicting variations in pdf shape than climatological error pdf
  - However, additional EPS skill tends to be relatively small initially.
  - It increases to max typically around $t \approx 4 \pm 2\,\mathrm{d}$.
  - Then, additional skill gradually decreases
- Similar results for T850, Z500, and also for MAM09, and for verification with ERA-Interim analyses (not shown)
- Initial skill increase consistent with fact that spread-error reliability improves with lead time
- Work in progress . . .
  - What should be expected from a good EPS system?
  - Can we get additional insight by using this technique to compare different ensemble configurations?
  - What can we learn from this for ensemble calibration techniques?

Evaluation of the pdf $p(x)$ of a continuous variable

- Two proper scores
  - Continuous Ranked Probability Score (CRPS)
  - Continuous Ignorance Score (CIgnS)
- Two reference forecasts are considered:
  - $N(\mathrm{CF}, \sigma^2_{\mathrm{err}}(\mathrm{CF}))$:
    Δscore between EPS and $N(\mathrm{CF}, \sigma^2_{\mathrm{err}})$ evaluates **all moments** of pdf
  - $N(\mathrm{EM}, \sigma^2_{\mathrm{err}}(\mathrm{EM}))$:
    Δscore between EPS and $N(\mathrm{EM}, \sigma^2_{\mathrm{err}})$ assesses **2nd and higher moments** of pdf

Evaluation of the pdf $p(x)$ of a continuous variable

- Two proper scores
  - Continuous Ranked Probability Score (CRPS)
  - Continuous Ignorance Score (CIgnS)
- Two reference forecasts are considered:
  - $N(\mathrm{CF}, \sigma^2_{\mathrm{err}}(\mathrm{CF}))$:
    Δscore between EPS and $N(\mathrm{CF}, \sigma^2_{\mathrm{err}})$ evaluates **all moments** of pdf
  - $N(\mathrm{EM}, \sigma^2_{\mathrm{err}}(\mathrm{EM}))$:
    Δscore between EPS and $N(\mathrm{EM}, \sigma^2_{\mathrm{err}})$ assesses **2nd and higher moments** of pdf
- What difference should be expected?
  - define two kinds of "perfect probabilistic forecast"
  - an analytical example
- Results for the operational ECMWF EPS

The Continuous Ranked Probability Score

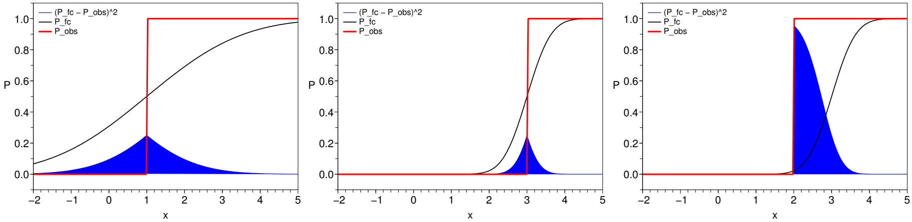
CRPS = Mean squared error of the cumulative distribution P_{fc}

$$\text{cdf of truth} \quad P_y(x) = P(y \leq x) = H(x-y) \tag{1}$$

$$\text{cdf of forecast} \quad P_{\mathrm{fc}}(x) = P(x_{\mathrm{fc}} \leq x) \tag{2}$$

$$\mathrm{CRPS} = \int \left(P_{\mathrm{fc}}(x) - P_y(x)\right)^2 \,\mathrm{d}x \tag{3}$$

$$= \int \mathrm{BS}_x \,\mathrm{d}x \tag{4}$$



The Continuous Ranked Probability Score

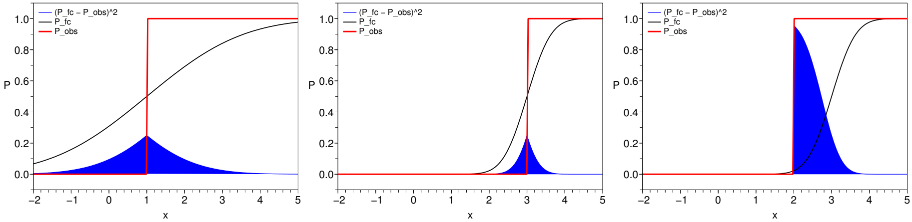
CRPS = Mean squared error of the cumulative distribution P_{fc}

$$\text{cdf of truth} \quad P_y(x) = P(y \le x) = H(x - y) \tag{1}$$

$$\text{cdf of forecast} \quad P_{\mathrm{fc}}(x) = P(x_{\mathrm{fc}} \le x) \tag{2}$$

$$\mathrm{CRPS} = \int (P_{\mathrm{fc}}(x) - P_y(x))^2 \, \mathrm{d}x \tag{3}$$

$$= \int \mathrm{BS}_x \, \mathrm{d}x \tag{4}$$



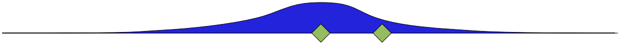
equal to Mean Absolute Error for a deterministic forecast

Continuous Ignorance Score

or Continuous Logarithmic Score

Let y denote truth and p the forecasted probability density

$$\mathrm{CIgnS} = -\log p(y)$$

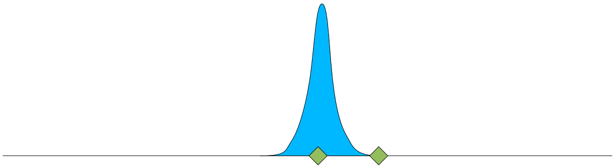


Continuous Ignorance Score

or Continuous Logarithmic Score

Let y denote truth and p the forecasted probability density

$$\mathrm{CIgnS} = -\log p(y)$$

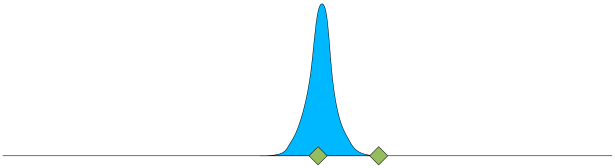


Continuous Ignorance Score

or Continuous Logarithmic Score

Let y denote truth and p the forecasted probability density

$$\mathrm{CIgnS} = -\log p(y)$$



For a Gaussian forecast $N(\mu, \sigma^2)$, we obtain

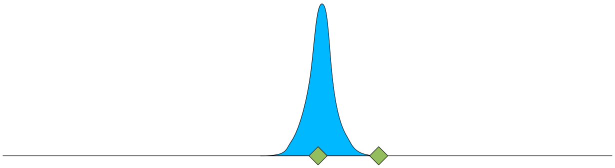
$$\mathrm{CIgnS} = \log(\sigma\sqrt{2\pi}) + \frac{(y-\mu)^2}{2\sigma^2}$$

Continuous Ignorance Score

or Continuous Logarithmic Score

Let y denote truth and p the forecasted probability density

$$\mathrm{CIgnS} = -\log p(y)$$



For a Gaussian forecast $N(\mu, \sigma^2)$, we obtain

$$\mathrm{CIgnS} = \log(\sigma\sqrt{2\pi}) + \frac{(y-\mu)^2}{2\sigma^2}$$

Mean squared error of reduced centred variable plus logarithmic penalty term for the spread (σ).

Perfect probabilistic forecasts

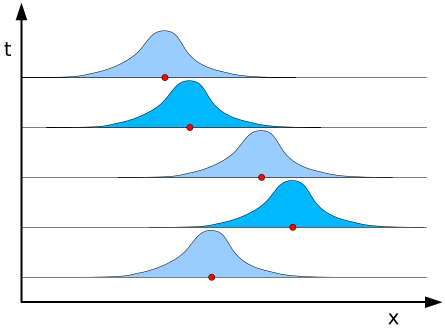
- Usually: skill score $= 0$ $\Rightarrow$ as good as climatology
  skill score $= 1$ $\Rightarrow$ perfect deterministic forecast
- We may still get closer to 1 but will *never* reach it!

Perfect probabilistic forecasts

- Usually: skill score $= 0 \quad \Rightarrow$ as good as climatology
  skill score $= 1 \quad \Rightarrow$ perfect deterministic forecast
- We may still get closer to 1 but will *never* reach it!
- Obs. and model uncertainties + perturbation growth characteristics of the atmosphere impose a lower limit on the forecast error variance $\sigma_f^2 > 0$.

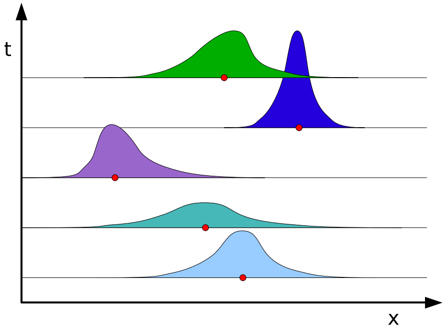
Perfect probabilistic forecasts

- Usually: skill score $= 0 \quad \Rightarrow$ as good as climatology
  skill score $= 1 \quad \Rightarrow$ perfect deterministic forecast
- We may still get closer to 1 but will *never* reach it!
- Obs. and model uncertainties + perturbation growth characteristics of the atmosphere impose a lower limit on the forecast error variance $\sigma_f^2 > 0$.
- Define a perfect probabilistic forecast under the constraint $\overline{v} \equiv \overline{\sigma_f^2} = \text{constant}$
  (consider a fixed lead time)



Perfect probabilistic forecasts

- Usually: skill score $= 0$ $\Rightarrow$ as good as climatology
  skill score $= 1$ $\Rightarrow$ perfect deterministic forecast
- We may still get closer to 1 but will *never* reach it!
- Obs. and model uncertainties + perturbation growth characteristics of the atmosphere impose a lower limit on the forecast error variance $\sigma_f^2 > 0$.
- Define a perfect probabilistic forecast under the constraint $\overline{v} \equiv \overline{\sigma_f^2} = \text{constant}$
  (consider a fixed lead time)



Levels of perfection

use label t to refer to different valid times of the forecast (lead time fixed)

Perfect dynamic forecast: Perfect flow- and data-dependent variations in pdf-shape

$$p_t(x) = p_d(x - \mu_t, t)$$

with p_d statistically consistent with error of the mean μ_t for each t,
given average variance $\mathbb{E}_t \int x^2 p_d(x, t) = \overline{v}$ and mean zero for each t.

Levels of perfection

use label t to refer to different valid times of the forecast (lead time fixed)

Perfect dynamic forecast: Perfect flow- and data-dependent variations in pdf-shape

$$p_t(x) = p_d(x - \mu_t, t)$$

with p_d statistically consistent with error of the mean μ_t for each t, given average variance $\mathbb{E}_t \int x^2 p_d(x,t) = \overline{v}$ and mean zero for each t.

Perfect static forecast: Constant (or seasonally varying) flow- and data-independent pdf-shape which is perfect:

$$p_t(x) = p_s(x - \mu_t)$$

with p_s statistically consistent with the error of the mean μ_t in the time-average sense, and $\int x p_s \, \mathrm{d}x = 0$, and $\int x^2 p_s \, \mathrm{d}x = \overline{v}$.

An idealized example with Gaussian distributions

Now, focus on variance prediction.

Let Ens. Mean error be a random variable distributed according to

$$p^*(x,t) = \frac{1}{\sigma(t)\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2(t)}\right)$$

- perfect dynamic forecast: issue $p_d = N(\mu_t, \sigma^2(t))$
- perfect static forecast: issue $p_s = N(\mu_t, \overline{\sigma^2})$ with $\overline{\sigma^2} = \mathbb{E}_t \sigma^2(t)$

An idealized example with Gaussian distributions

Now, focus on variance prediction.

Let Ens. Mean error be a random variable distributed according to

$$p^*(x,t) = \frac{1}{\sigma(t)\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2(t)})$$

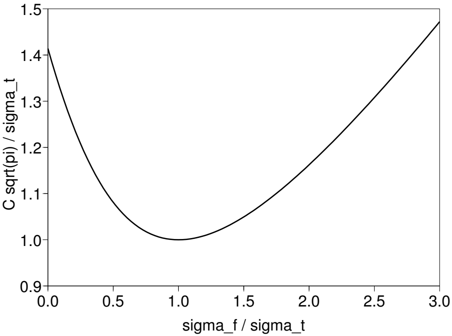
- perfect dynamic forecast: issue $p_d = N(\mu_t, \sigma^2(t))$
- perfect static forecast: issue $p_s = N(\mu_t, \overline{\sigma^2})$ with $\overline{\sigma^2} = \mathbb{E}_t \sigma^2(t)$

What is the difference in probabilistic scores (CRPS, CIgnS) between the perfect dynamic forecast and the perfect static forecast?

Expected value of CRPS

Let y denote the true value of the EM error. We see that the expected value of the CRPS is

$$\mathbb{E}_y \, \mathrm{CRPS}(N(0,\sigma_f^2), y) = \frac{\sigma_t}{\sqrt{\pi}} \left[-\frac{\sigma_f}{\sigma_t} + \sqrt{2 + 2\sigma_f^2/\sigma_t^2} \right]$$

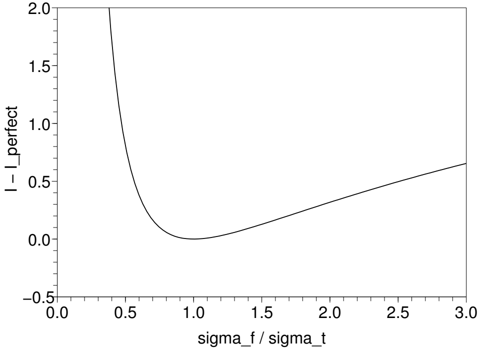


The CRPS has a minimum value at $\sigma_f = \sigma_t$. This is not surprising as the CRPS is a proper score.

Expected value of CIgnS

Again let y denote the true value of the EM error. The expected value of the Continuous Ignorance Score is

$$\mathbb{E}_y \, \mathrm{CIgnS}(N(0,\sigma_f^2), y) = \frac{1}{2}\left[\ln(2\pi\sigma_f^2) + (\sigma_t/\sigma_f)^2\right]$$

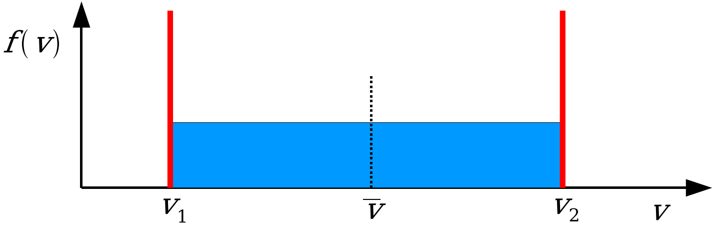


The minimum is again at $\sigma_f = \sigma_t$; CIgnS is proper!

Two particular distributions of variance

Let $v = \sigma^2$ denote the variance

- continuous uniform distribution: $v \sim U(v_1, v_2)$
- discrete uniform distribution: $v \sim \frac{1}{2}\delta(v - v_1) + \frac{1}{2}\delta(v - v_2)$

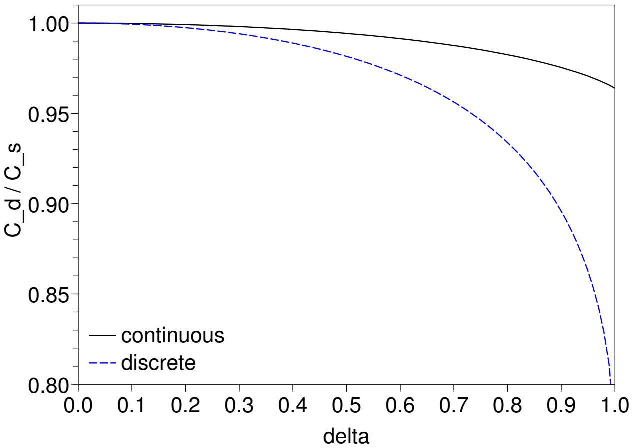


Introduce dimensionless parameter

$$\delta = \frac{v_2 - v_1}{2\overline{v}} \in [0, 1]$$

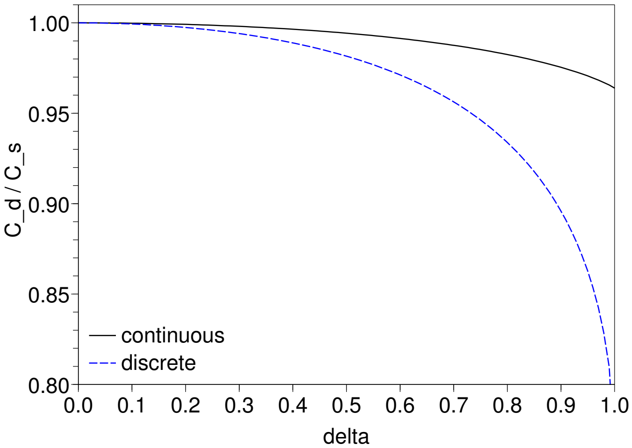
Expected CRPS for uniform variance distributions

CRPS ratio: dynamic forecast / static forecast



Expected CRPS for uniform variance distributions

CRPS ratio: dynamic forecast / static forecast

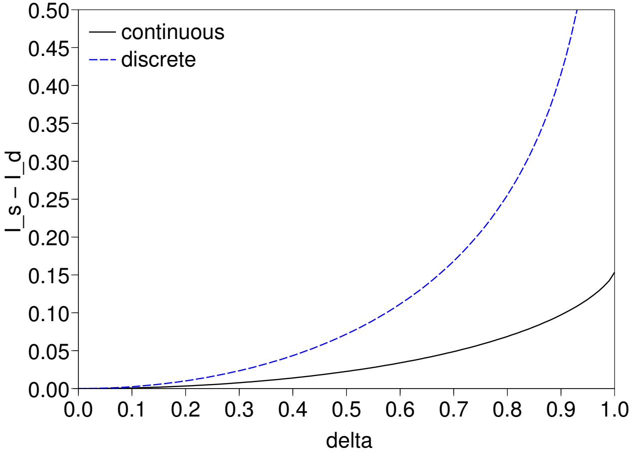


for 48-hour error doubling:

3% (20%) reduction in CRPS $\Longrightarrow$ 2-hour (13-hour) gain in lead time

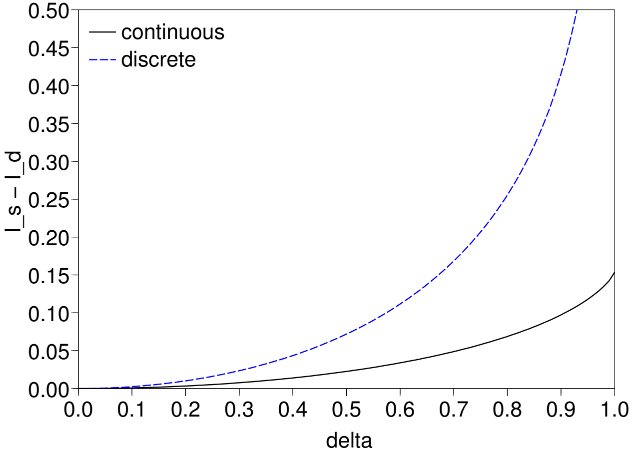
Expected CIgnS for uniform variance distributions

CIgnS(static forecast) − CIgnS(dynamic forecast)



Expected CIgnS for uniform variance distributions

CIgnS(static forecast) − CIgnS(dynamic forecast)



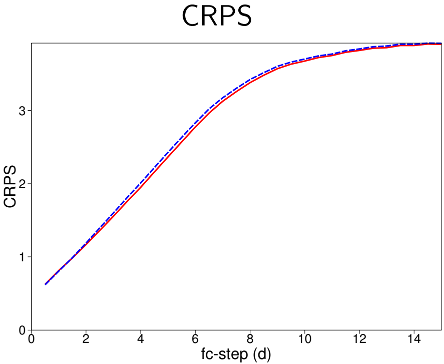
for 48-hour error doubling:

reduction of CIgnS by 0.15 (0.7) ⟹ 10-hour (48-hour) gain in lead time

Dressed ens. mean forecast: v 850 hPa, 35°–65°N, DJF09

EPS (red solid line) — raw prob. for CRPS; Gaussian for CIgnS

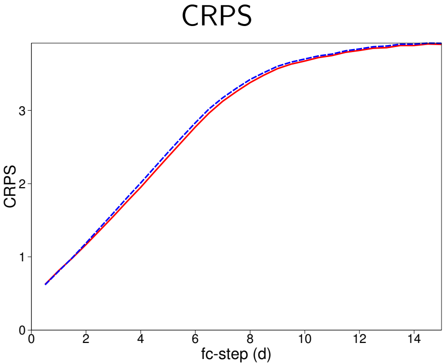
$N(\mathrm{EM}, \sigma^2_{\mathrm{err}}(\mathrm{EM}))$ (blue dashed line) — σ_{err} estimated from reforecasts

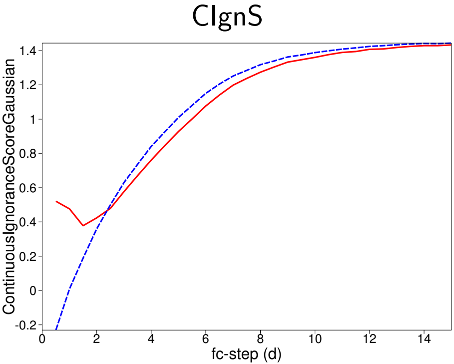


Dressed ens. mean forecast: v 850 hPa, 35°–65°N, DJF09

EPS — raw prob. for CRPS; Gaussian for CIgnS

$N(\mathrm{EM}, \sigma^2_{\mathrm{err}}(\mathrm{EM}))$ — σ_{err} estimated from reforecasts

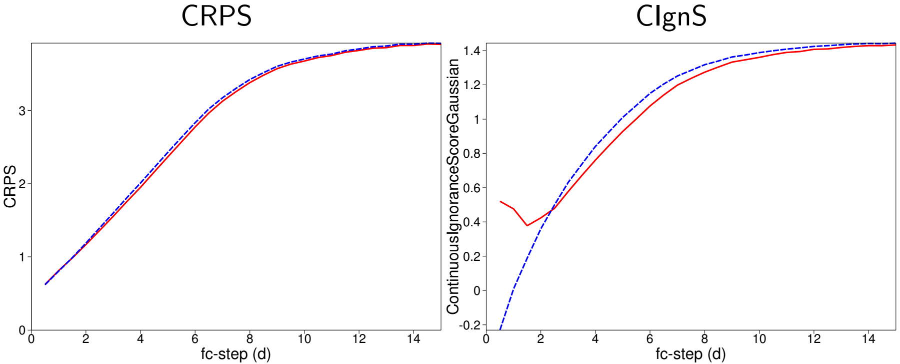




Dressed ens. mean forecast: v 850 hPa, 35°–65°N, DJF09

EPS — raw prob. for CRPS; Gaussian for CIgnS

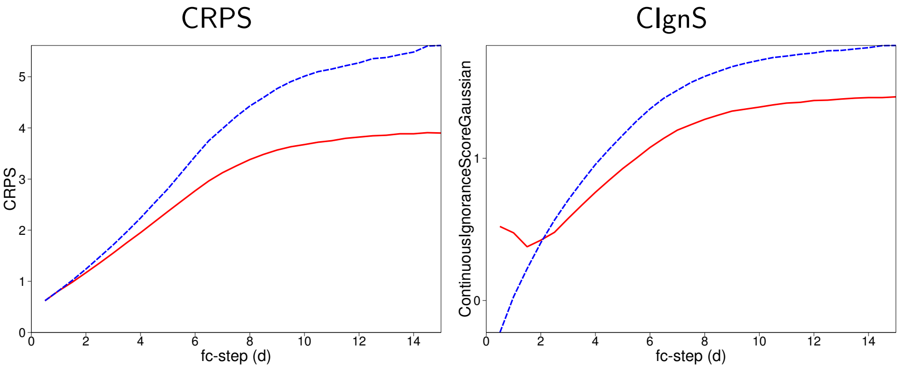
$N(\mathrm{EM}, \sigma^2_{\mathrm{err}}(\mathrm{EM}))$ — σ_{err} estimated from reforecasts



- T850, Z500, qualitatively similar but . . .
- Deficiencies in the short-range can be addressed via calibration (to a certain extent).

Dressed control forecast: v 850 hPa, 35°–65°N, DJF09

- EPS — raw prob. for CRPS; Gaussian for CIgnS
- $N(\mathrm{CF}, \sigma^2_{\mathrm{err}}(\mathrm{CF}))$ — σ_{err} estimated from reforecasts



Diagnosis & Numerical Experimentation

Initial uncertainty

- Deeper understanding from applying diagnostic techniques to clean numerical experimentation designed to answer specific questions

Diagnosis & Numerical Experimentation

Initial uncertainty

- Deeper understanding from applying diagnostic techniques to clean numerical experimentation designed to answer specific questions
- In the early ranges, say up to day 2, EM dressed with a climatological error distribution as good as or better than EPS.

Diagnosis & Numerical Experimentation

Initial uncertainty

- Deeper understanding from applying diagnostic techniques to clean numerical experimentation designed to answer specific questions
- In the early ranges, say up to day 2, EM dressed with a climatological error distribution as good as or better than EPS.
- If CF/EM + past errors provide skilful probabilistic forecasts, then one may ask whether past errors might be a successful EPS perturbation strategy

Flow-independent perturbations

- Mureau, Molteni & Palmer (1993)
  - initial perturbations based on 6-hour forecast errors from past 30 days & Gram-Schmidt-orthonormalisation
  - assimilation OI
  - model T63

Flow-independent perturbations

- Mureau, Molteni & Palmer (1993)
  - initial perturbations based on 6-hour forecast errors from past 30 days & Gram-Schmidt-orthonormalisation
  - assimilation OI
  - model T63
  - conclusion: SV perturbations are superior
- Magnusson, Nycander & Källén (2008): flow-independent perts. constructed from scaled differences of randomly picked atmospheric states.

Flow-independent perturbations

- Mureau, Molteni & Palmer (1993)
  - initial perturbations based on 6-hour forecast errors from past 30 days & Gram-Schmidt-orthonormalisation
  - assimilation OI
  - model T63
  - conclusion: SV perturbations are superior
- Magnusson, Nycander & Källén (2008): flow-independent perts. constructed from scaled differences of randomly picked atmospheric states. Initially quite overdispersive in Z500, but skill close to ensemble using operational SV perturbations.

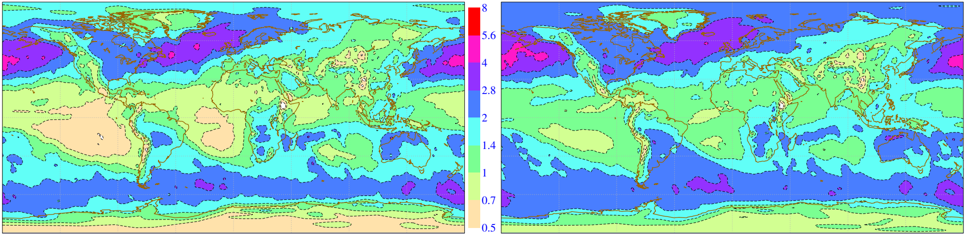
Flow-independent perturbations

- Mureau, Molteni & Palmer (1993)
  - initial perturbations based on 6-hour forecast errors from past 30 days & Gram-Schmidt-orthonormalisation
  - assimilation OI
  - model T63
  - conclusion: SV perturbations are superior
- Magnusson, Nycander & Källén (2008): flow-independent perts. constructed from scaled differences of randomly picked atmospheric states. Initially quite overdispersive in Z500, but skill close to ensemble using operational SV perturbations.
- Here: use random sample from past 24-hour forecast errors as initial perturbations (advantage: characteristics of short-range fc errors are closer to those of analysis errors than scaled differences of full fields)

Time mean spread vs. RMSE of Ens. mean

Meridional wind component ($m\,s^{-1}$) at 850 hPa, t=48 h

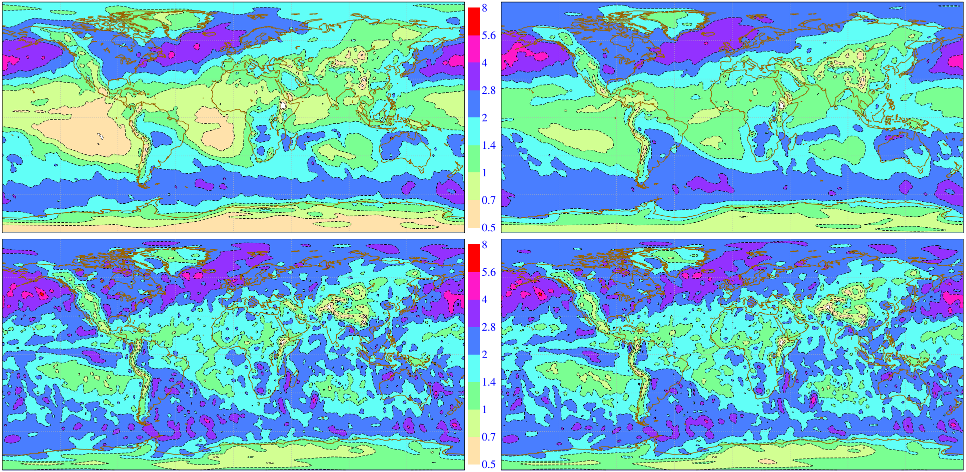
singular vector init. perts. | 24-hour fc. error init. perts.



Time mean spread vs. RMSE of Ens. mean

Meridional wind component (m s^{-1}) at 850 hPa, t=48 h

singular vector init. perts. | 24-hour fc. error init. perts.

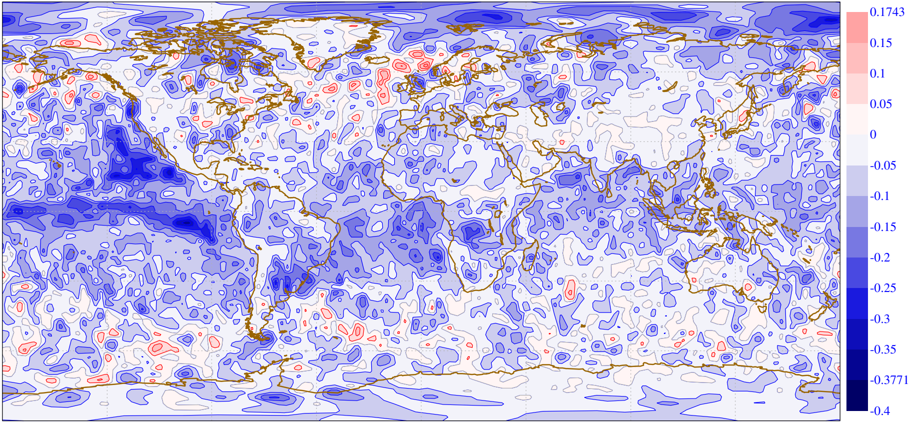


top: ens. stdev.; **bottom:** ens. mean RMS error; 50 cases: 23 Nov '07–29 Feb '08

T_L255, 32r3, unscaled 24-hour FCEs

CRPS difference: FCE − SV

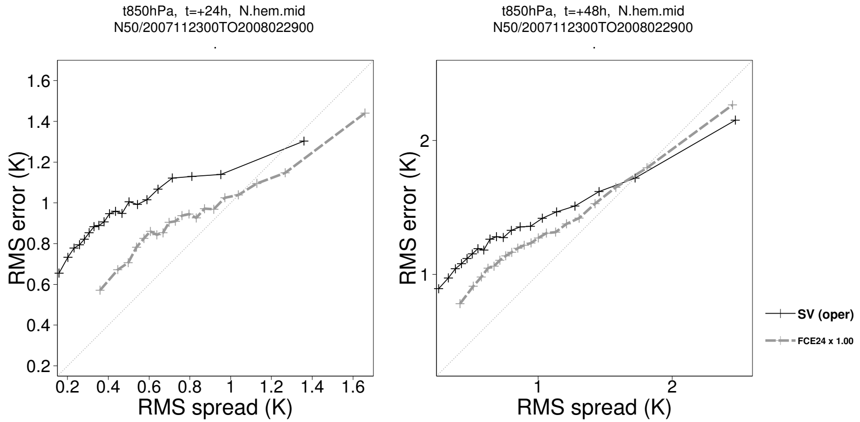
Meridional wind component ($m\,s^{-1}$) at 850 hPa, t=48 h



- CRPS (Continuous Ranked Probability Score ≡ mean squared error of the cumulative distribution)
- Blue means EPS based on short-range forecast errors is more skilful.
- 50 cases: 23 Nov 2007 – 29 Feb 2008

Spread-reliability

850 hPa temperature, 35°–65°N



Conclusions

- Comparison of Spread and EM-error continues to be an essential tool
  - Have not achieved a well tuned system for all variables and regions.
  - Achieving a reliable distribution of spread in space and time in the early forecast ranges is one of the major challenges for the future.

Conclusions

- Comparison of Spread and EM-error continues to be an essential tool
  - Have not achieved a well tuned system for all variables and regions.
  - Achieving a reliable distribution of spread in space and time in the early forecast ranges is one of the major challenges for the future.
- Moment-based decomposition of the ensemble skill has been explored
  - Skill of Ens. mean + skill of pdf of Ens. mean errors
  - Continuous Ignorance Score appears better suited than CRPS to evaluate flow- and data-dependent variations in spread.

Conclusions

- Comparison of Spread and EM-error continues to be an essential tool
  - Have not achieved a well tuned system for all variables and regions.
  - Achieving a reliable distribution of spread in space and time in the early forecast ranges is one of the major challenges for the future.
- Moment-based decomposition of the ensemble skill has been explored
  - Skill of Ens. mean + skill of pdf of Ens. mean errors
  - Continuous Ignorance Score appears better suited than CRPS to evaluate flow- and data-dependent variations in spread.
- Initial perturbations based on past short-range forecast errors
  - $\rightarrow$ challenging benchmark for flow-dependent initial perturbations (e.g. singular vectors).
  - Rare locally large perturbations that are inconsistent with the flow may be an obstacle for operational implementation of the benchmark system.
  - Further diagnostic work in this area is expected to help the development of ensemble prediction system with improved flow-dependent variations of the pdf (in particular in the earlier forecast ranges).

Issues to verify “initial uncertainty”

- truth not available
- short-range fc error correlated with analysis error $\rightarrow$ require obs (or perhaps an independent analysis)
- obs uncertainty/analysis uncertainty needs to be accounted for if ensemble spread is smaller than or of similar magnitude as the obs/an uncertainty.

Multivariate verification

- Why? Ensemble of assimilations should provide a varying background error covariance to the assimilation system. One should attempt to verify the flow-dependent **co**variances.
- There are user applications that will be dependent on the joint pdf of several variables (e.g. distributed in space or several variables). Implications for suitable ensemble size can be different from univariate case.