Ensemble size: How suboptimal is less than infinity?

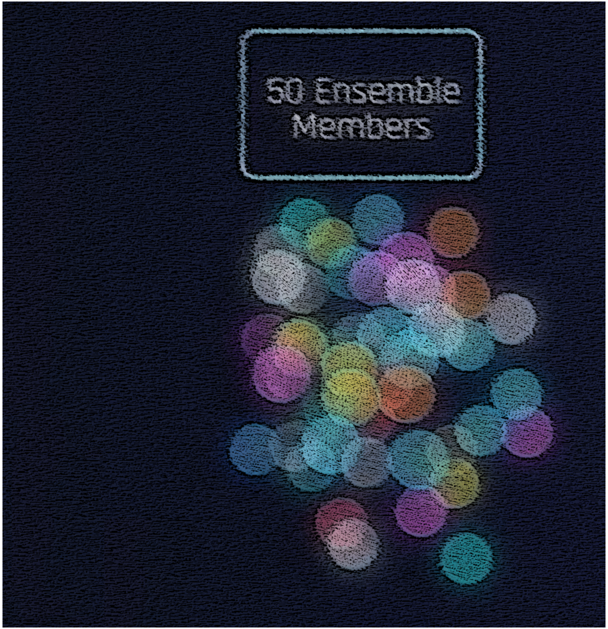
Martin Leutbecher

13 September 2017

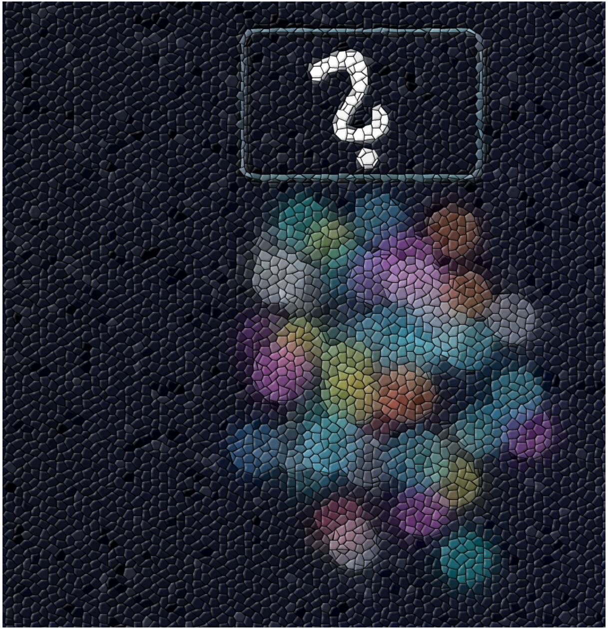
Acknowledgements: Zied Ben Bouallègue, Chris Ferro, Sarah-Jane Lock, David Richardson

Ensemble size at ECMWF

Ensemble size at ECMWF



Ensemble size at ECMWF



50 member since Dec 1996

Why 50?

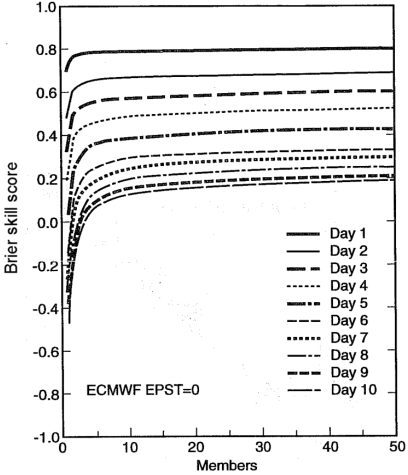
Ensemble size at ECMWF

50 member since Dec 1996

Why 50?

Are the benefits of more than 50 members marginal?

Talagrand, Vautard and Strauss (1997)



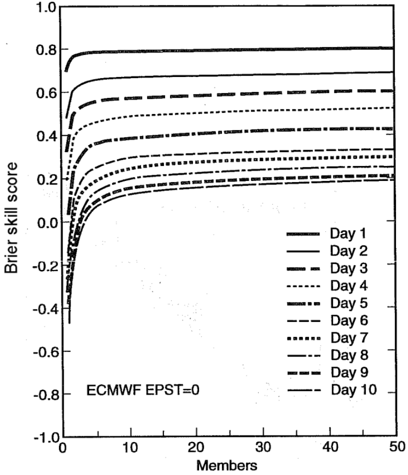
(adapted from their Fig. 4.5)

4.2.3 *Dependence on the size of forecast ensembles*

One particularly interesting question is whether one should continue increasing the size of the ensembles or rather concentrate efforts on other points. Figure 4.5 attempts to address this issue. We display the global BSS values as a function of the number of members N, for the median threshold (τ = 0K) and the extreme threshold (τ = 8K). One argument for the extension of the ensemble size is the better estimation of probabilities of extreme events. We should therefore see in Figure 4.5 a larger sensitivity to N for the threshold τ = 8K than for the threshold τ = 0K. Such is not the case. It is to be noticed that convergence is actually reached quite quickly at all lead times, for, say, N = 20-30.

...

Talagrand, Vautard and Strauss (1997)



(adapted from their Fig. 4.5)

4.2.3 *Dependence on the size of forecast ensembles*

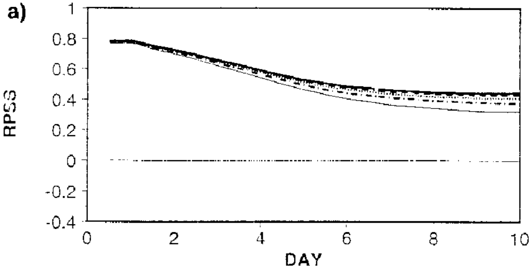
One particularly interesting question is whether one should continue increasing the size of the ensembles or rather concentrate efforts on other points. Figure 4.5 attempts to address this issue. We display the global BSS values as a function of the number of members N, for the median threshold (τ = 0K) and the extreme threshold (τ = 8K). One argument for the extension of the ensemble size is the better estimation of probabilities of extreme events. We should therefore see in Figure 4.5 a larger sensitivity to N for the threshold τ = 8K than for the threshold τ = 0K. Such is not the case. It is to be noticed that convergence is actually reached quite quickly at all lead times, for, say, N = 20-30. . . .

According to Talagrand et al (1997) not more than 30 members are needed.

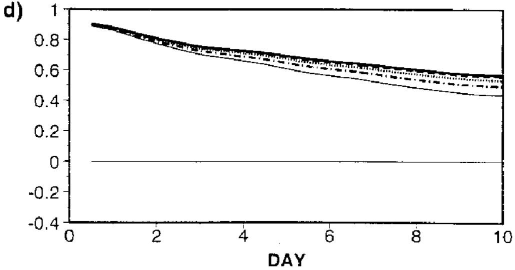
Buizza and Palmer (1998)

Comparison of 2, 4, 8, 16, 32 members

Verified against analysis



Verified against member



(adapted from their Fig. 11) Z500 in NH; T63L19 model, initial uncertainty represented with singular vectors, no representation of model uncertainty

Careful conclusions that do not rule out increases in skill beyond 32 members.

Buizza et al. (1998)

Impact of resolution and ensemble size

TABLE 2. CHARACTERISTICS OF THE ENSEMBLE PREDICTION SYSTEM (EPS) CONFIGURATIONS TESTED

EPS configuration	Member size	Forecast resolution	Singular vectors' resolution
32*T63	32	T63L19	T42L19
128*T63	128	T63L19	T42L19
32*T106	32	T106L31	T42L19
32*T106SV31	32	T106L31	T42L31
50*T106SV31	50	T106L31	T42L31

TABLE 9. BRIER SKILL SCORE FOR PROBABILITY PREDICTION OF PRECIPITATION AMOUNTS OF 1 AND 10 MM DAY^{-1}, OVER THE NORTHERN HEMISPHERE, AT FORECAST-DAYS 5 AND 7

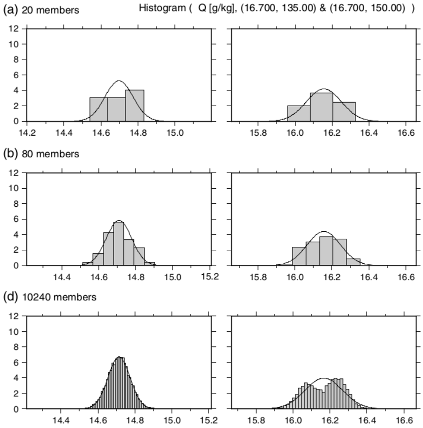
	Forecast-day 5		Forecast-day 7	
Configuration	1 mm day^{-1}	10 mm day^{-1}	1 mm day^{-1}	10 mm day^{-1}
32*T63	0.286	0.066	0.201	0.009
32*T106	0.286	0.095	0.219	0.078
32*T106SV31	0.285	0.097	0.219	0.078
50*T106SV31	0.298	**0.104**	0.230	**0.091**
128*T63	**0.299**	0.087	**0.238**	0.049

For each precipitation amount, bold figures identify the most skilful results.

In December 1996, resolution was increased from T63 (quadratic grid) to TL159 (linear grid) and ensemble size was increased from 32 to 50 members.

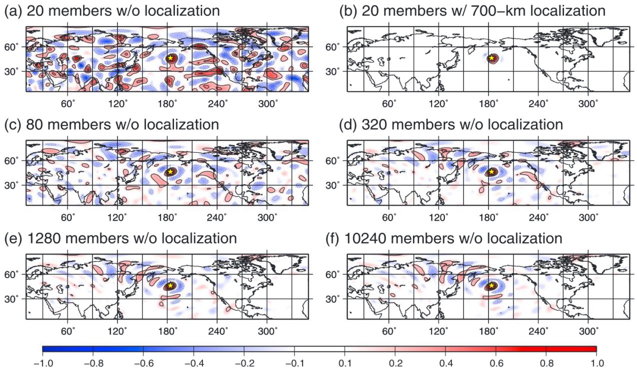
Miyoshi et al (2014)

identical twin EnKF using SPEEDY model



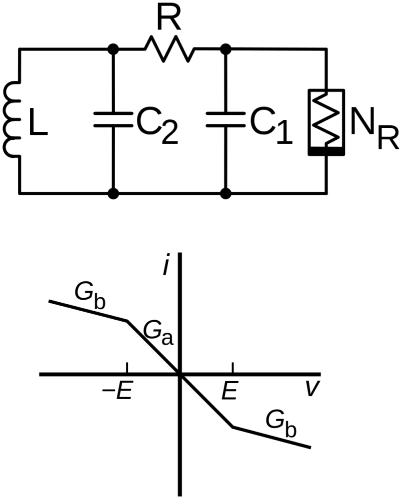
(from their Fig. 3)

Horizontal correlations of mid-tropospheric specific humidity with "yellow star" location



(from their Fig. 4)

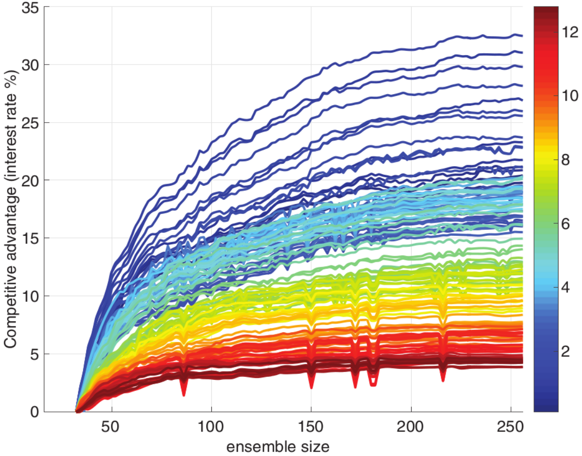
Chua's circuit



(from Wikipedia)

Machete and Smith (2016)

ensemble forecasts of electronic circuit

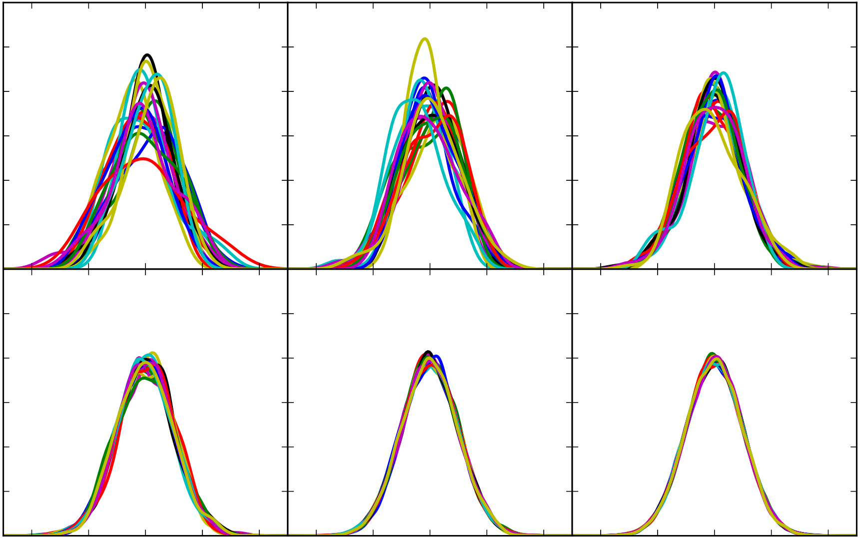


(from their Fig. 12; colour corresponds to lead time; competitive advantage is similar to a probabilistic skill score)

ECMWF

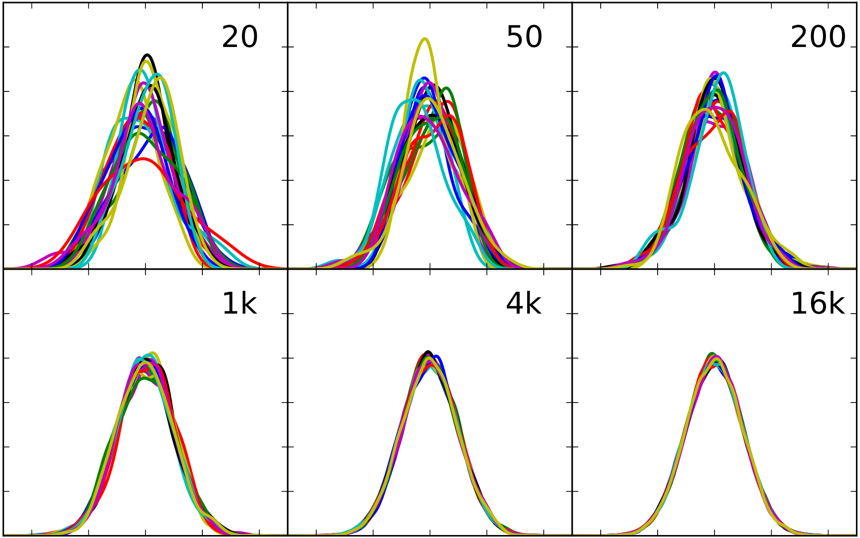
Guess the ensemble size

i.i.d. members; pdfs for 20 realisations; ensemble size fixed in each panel



Guess the ensemble size

i.i.d. members; pdfs for 20 realisations; ensemble size fixed in each panel

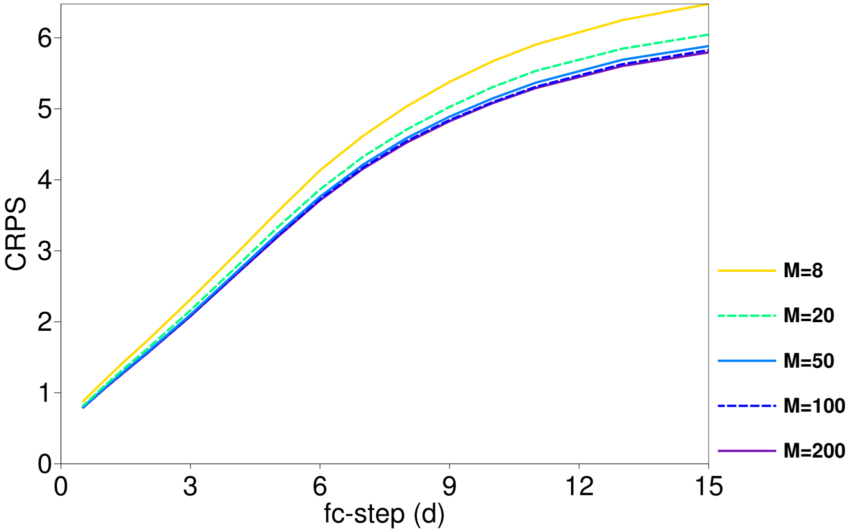


Ensemble size at ECMWF

Experiments with IFS ensembles

- IFS cycle 41r2
- as operational ensemble but lower resolution: TCo399
- 200 members
- June-July-August 2016 (92 cases)
- probabilistic skill evaluated with continuous ranked probability score (CRPS): mean squared error of cumulative distribution

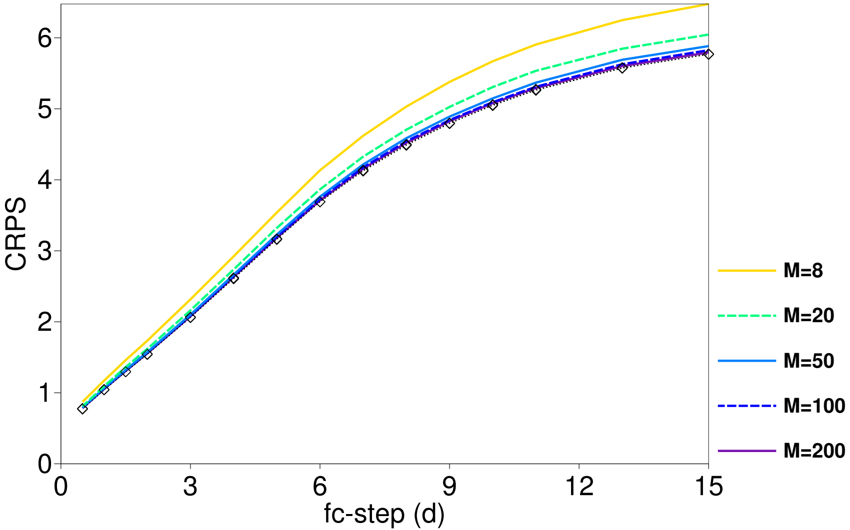
Impact of ensemble size on CRPS



200 hPa zonal wind in northern extra-tropics

Impact of ensemble size on CRPS

Predictions of CRPS for infinite ensemble size ···◇···◇···◇···

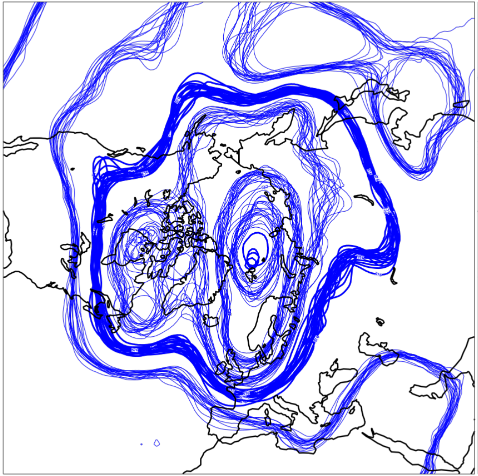


200 hPa zonal wind in northern extra-tropics

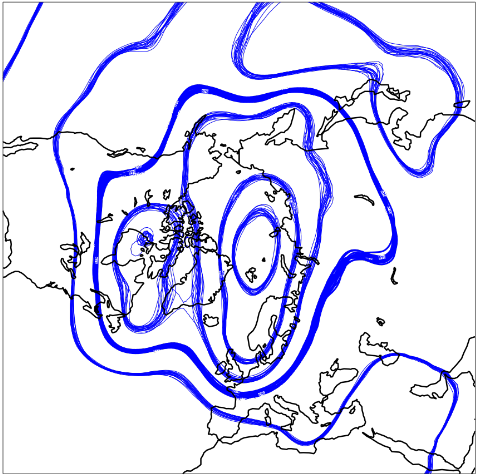
Where does increased skill come from?

Sampling uncertainty of Z500 ensemble mean at D7

8 member



50 member



CRPS and ensemble size: What to expect?

Kernel representation of CRPS

- kernel representation of CRPS

$$\mathrm{CRPS}(x_j, y) = \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|$$

- With exchangeability of members, the expected CRPS is

$$\mathbb{E}_x \mathrm{CRPS}(x_j, y) = \mathbb{E}_x |x - y| - \frac{M-1}{2M}\mathbb{E}_{x,x'}|x - x'|$$

- For an infinite size ensemble we get

$$\mathbb{E}_x \mathrm{CRPS}(x_j, y) = \mathbb{E}_x |x - y| - \frac{1}{2}\mathbb{E}_{x,x'}|x - x'|$$

How can CRPS for infinite ensemble size be predicted with a finite ensemble?

- The fair CRPS is a modified version of the CRPS that removes the bias in the score due to the finite ensemble size (see Chris Ferro's talk)
- From the kernel representation, one can see easily that the CRPS for infinite ensemble size is obtained by the estimator

$$\mathrm{CRPS}^*(x_j, y) = \mathrm{CRPS}(x_j, y) - \frac{1}{2M^2(M-1)} \sum_{j=1}^{M} \sum_{k=1}^{M} |x_j - x_k|$$

- The correction term is a measure of ensemble spread.

Analytic result for statistically consistent ensembles

- When members are statistically consistent (iid) draws from same distribution as observation (perfectly reliable ensemble), the CRPS for an m-member ensemble satisfies

$$\mathrm{CRPS}_M = \left(1 - \frac{M-1}{2M}\right) \mathbb{E}\left|x - x'\right| = \left(1 + \frac{1}{M}\right) \mathrm{CRPS}_\infty$$

- Eqns. (8) and (9) in Richardson (2001) show that the Brier score also satisfies $\mathrm{BS}_M = (1 + M^{-1})\,\mathrm{BS}_\infty$.

Analytic result for statistically consistent ensembles

- When members are statistically consistent (iid) draws from same distribution as observation (perfectly reliable ensemble), the CRPS for an m-member ensemble satisfies

$$\mathrm{CRPS}_M = \left(1 - \frac{M-1}{2M}\right) \mathbb{E}\left|x - x'\right| = \left(1 + \frac{1}{M}\right) \mathrm{CRPS}_\infty$$

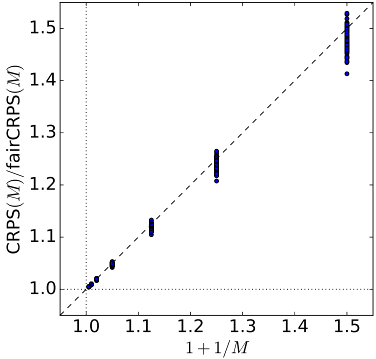
- Eqns. (8) and (9) in Richardson (2001) show that the Brier score also satisfies $\mathrm{BS}_M = (1 + M^{-1})\,\mathrm{BS}_\infty$.
- Extreme events? Relationship for BS implies that for any weighting in the twCRPS (Gneiting and Ranjan, 2011) we also have

$$\mathrm{twCRPS}_M = \left(1 + \frac{1}{M}\right) \mathrm{twCRPS}_\infty$$

Actual convergence with ensemble size

from right to left

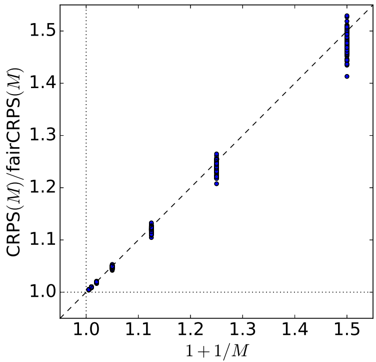
2, 4, 8, 20, 50, 100 and 200 members



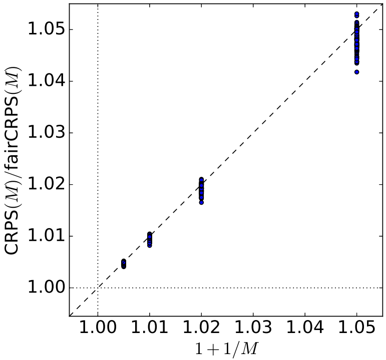
- Data from 200 member TCo399 IFS experiment, JJA2016
- 120 data points for each ensemble size
- 15 lead times $\times$ 4 variables (z500, T850, u850, u200) $\times$ 2 regions (NH and SH extratropics)
- 50 and 200 members are 2% and 0.5% worse than ∞, respectively

Actual convergence with ensemble size

from right to left
2, 4, 8, 20, 50, 100 and 200 members

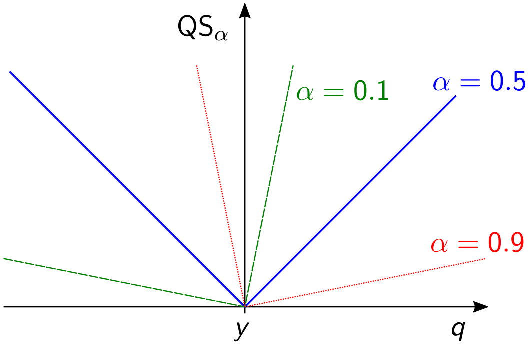


zoom
20, 50, 100 and 200 members



- Data from 200 member TCo399 IFS experiment, JJA2016
- 120 data points for each ensemble size
- 15 lead times $\times$ 4 variables (z500, T850, u850, u200) $\times$ 2 regions (NH and SH extratropics)
- 50 and 200 members are 2% and 0.5% worse than ∞, respectively

Quantile score and CRPS



$$\mathrm{QS}_\alpha(q, y) = 2\left(\mathbb{I}\{y < q\} - \alpha\right)(q - y)$$

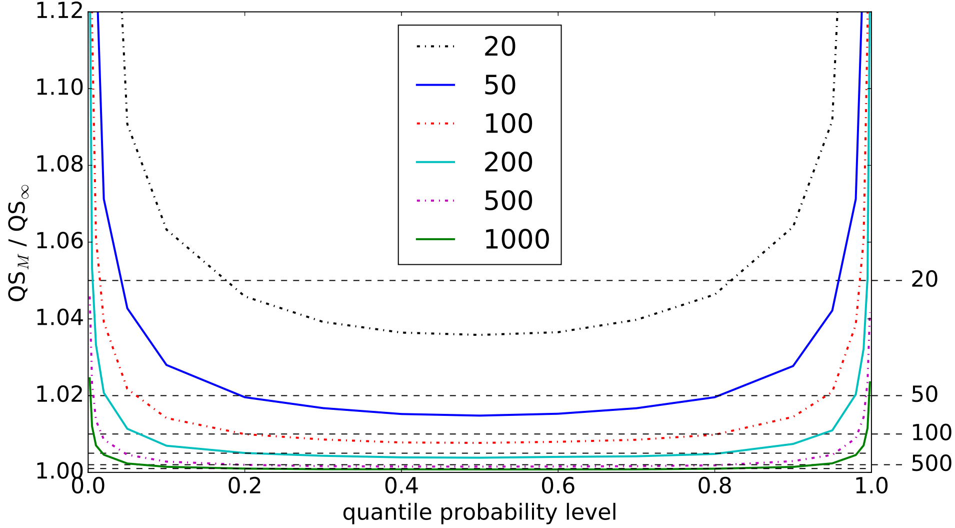
with indicator function $\mathbb{I}(\text{true}) = 1$ and $\mathbb{I}(\text{false}) = 0$, quantile q and observation y; $\alpha \in (0, 1)$ denotes the probability level

$$\mathrm{CRPS}(F, y) = \int_0^1 \mathrm{QS}_\alpha\left(F^{-1}(\alpha), y\right)\, \mathrm{d}\alpha$$

where the quantile q for cumulative distribution F is $F^{-1}(\alpha)$

Quantile score for a standard Gaussian

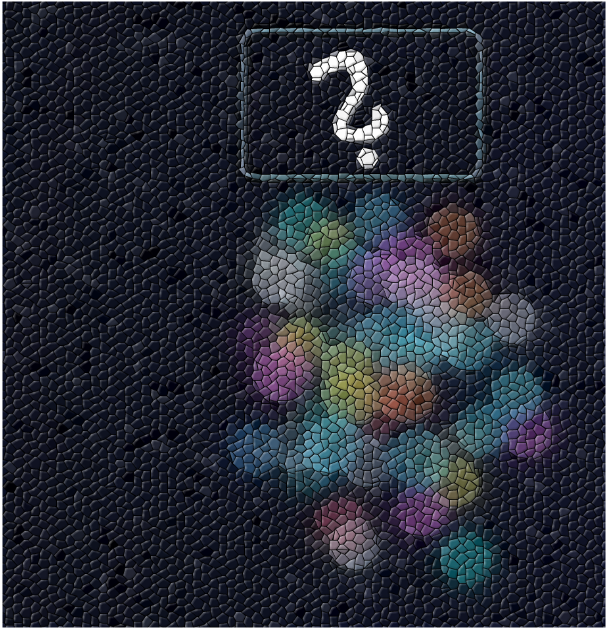
Simulations with $M = 20$ to 1000 members



For QS of $q_{.98}$, 50 and 200 members are 7% and 2% worse than ∞, respectively.

ECMWF

Ensemble size at ECMWF



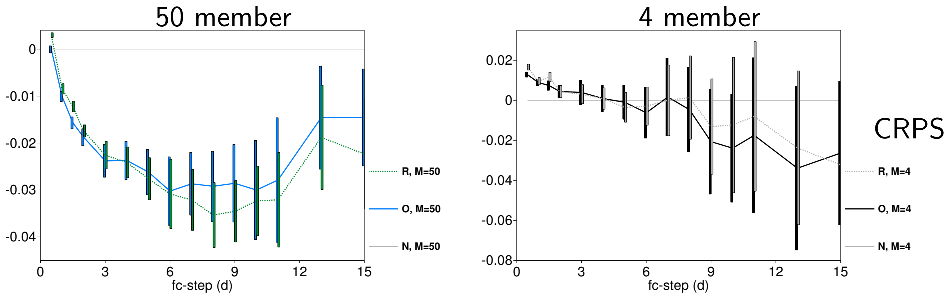
Research and development

What is a good ensemble size?

- Large ensemble size can delay progress in R&D
- It would be most efficient to use the smallest ensemble size that is sufficient to estimate impact for operational ensemble size
- Using proper scores with small ensembles can mislead though

Ensemble configurations R, O and N

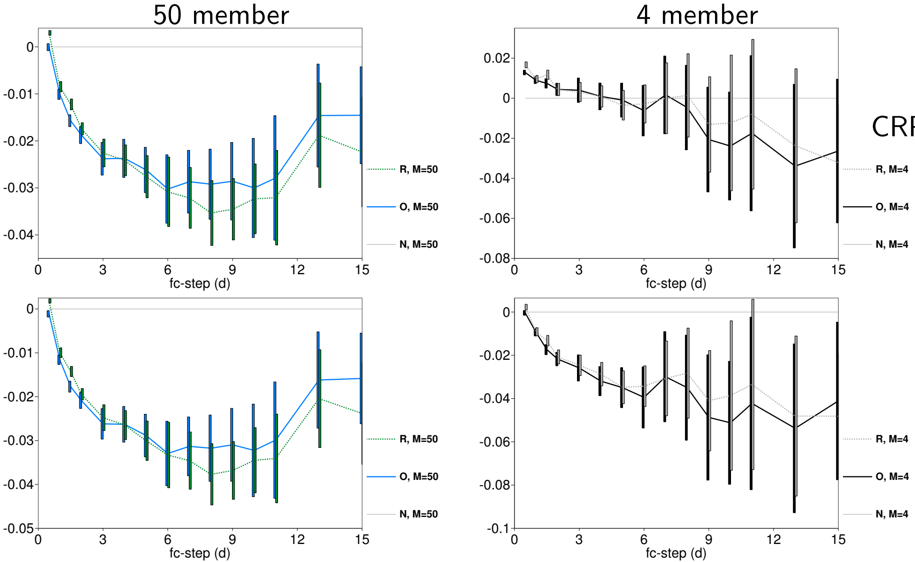
ΔCRPS for 850 hPa temperature in northern extratropics



CRPS… − CRPS_N

Ensemble configurations R, O and N

ΔCRPS for 850 hPa temperature in northern extratropics

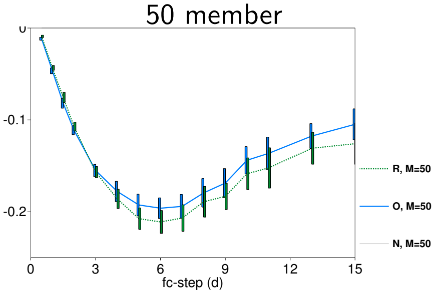


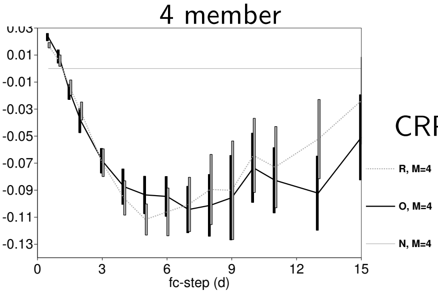
$\mathrm{CRPS}_{\ldots} - \mathrm{CRPS}_N$

fair CRPS

Ensemble configurations R, O and N

ΔCRPS for 850 hPa zonal wind in tropics

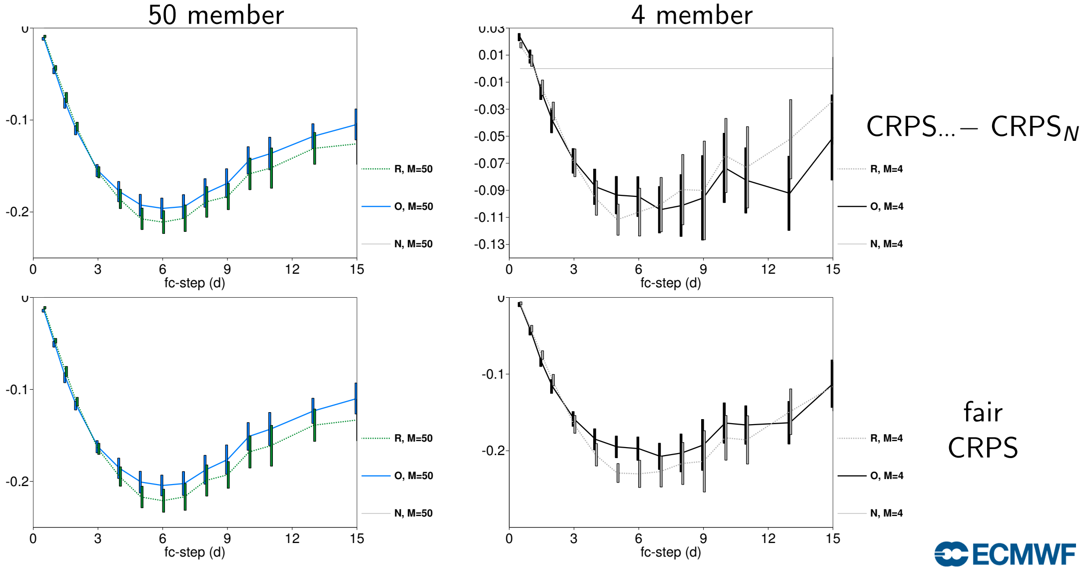




$\mathrm{CRPS}_{\ldots} - \mathrm{CRPS}_N$

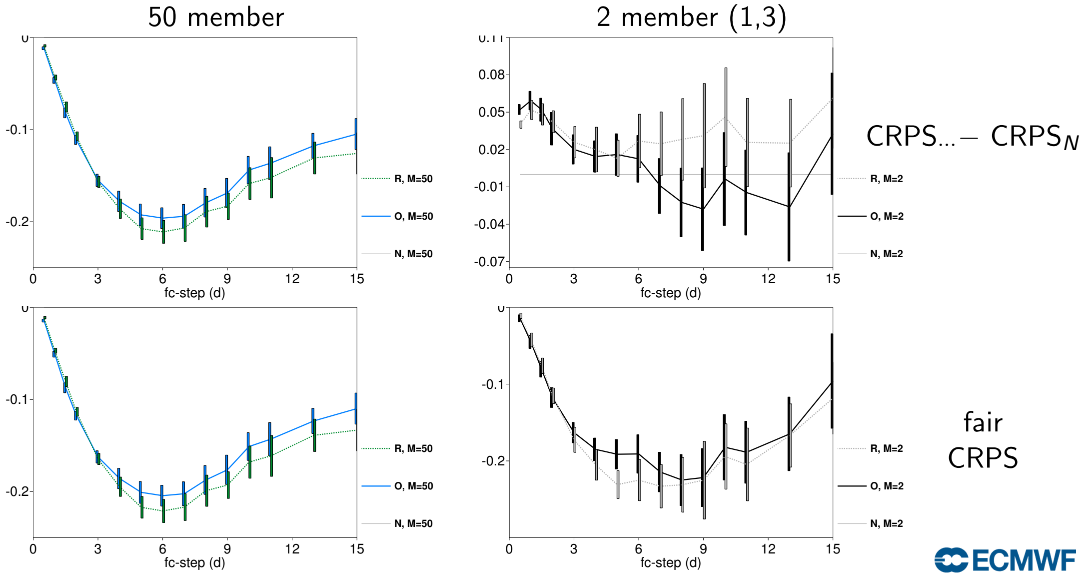
Ensemble configurations R, O and N

ΔCRPS for 850 hPa zonal wind in tropics



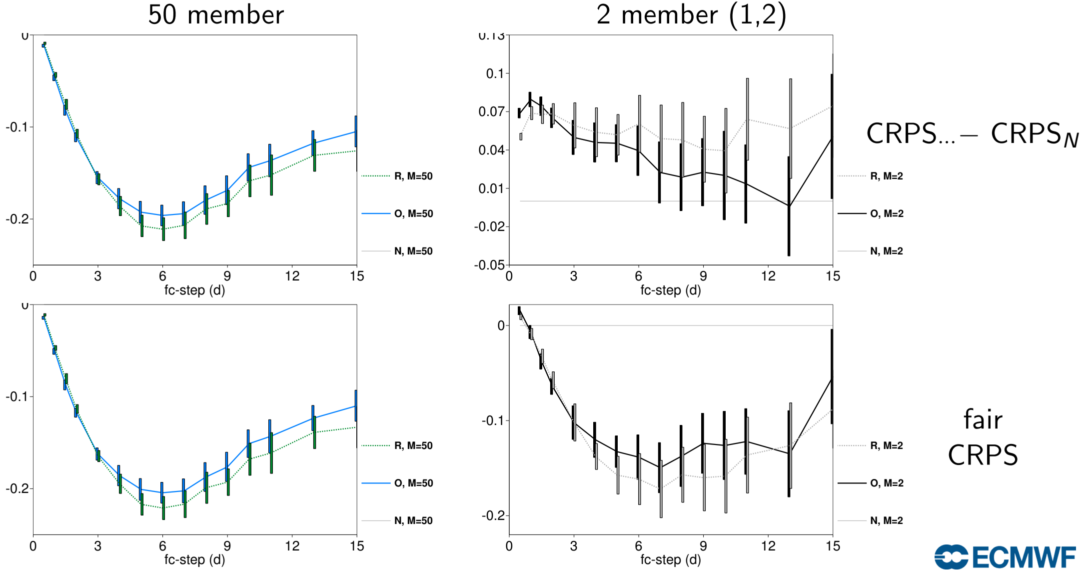
Ensemble configurations R, O and N

ΔCRPS for 850 hPa zonal wind in tropics



Ensemble configurations R, O and N

ΔCRPS for 850 hPa zonal wind in tropics



Small ensemble sizes

- Can be used for R&D if evaluation uses fair scores
- Can be used in reforecasts for estimating skill
- Applicability of fair scores is linked to ensemble generation
- Current ensemble generation at ECMWF not fully consistent with exchangeability required for fair scores
- Benefits for R&D
  - faster turnaround time
  - more configurations can be explored
  - scope for increasing statistical significance by using less members but more start dates

Conclusions

How suboptimal is less than infinity?

Three possible answers:

Conclusions

How suboptimal is less than infinity?

Three possible answers:

- A bit or maybe a lot, tell me the score and your ensemble size . . .

Conclusions

How suboptimal is less than infinity?

Three possible answers:

- A bit or maybe a lot, tell me the score and your ensemble size . . .
- **Operational ensemble forecasts:** 50 members are too few — let's increase the ensemble size to . . .

Conclusions

How suboptimal is less than infinity?

Three possible answers:

- A bit or maybe a lot, tell me the score and your ensemble size . . .
- **Operational ensemble forecasts:** 50 members are too few — let's increase the ensemble size to . . .
- **Research & Development:** Small ensembles are highly efficient.
  Two to four members may be enough for standard evaluations (provided exchangeability in the ensemble generation and use of fair scores)

Discussion

- How can we increase ensemble size when we need to increase resolution too?
- Different users will have different needs, how to obtain a good compromise for all of them?
- How to increase ensemble size in a computationally efficient way for all forecast ranges from medium-range to extended-range?
- What is an adequate ensemble size for the reforecasts?
- Which other proper scores permit the construction of an associated fair score?