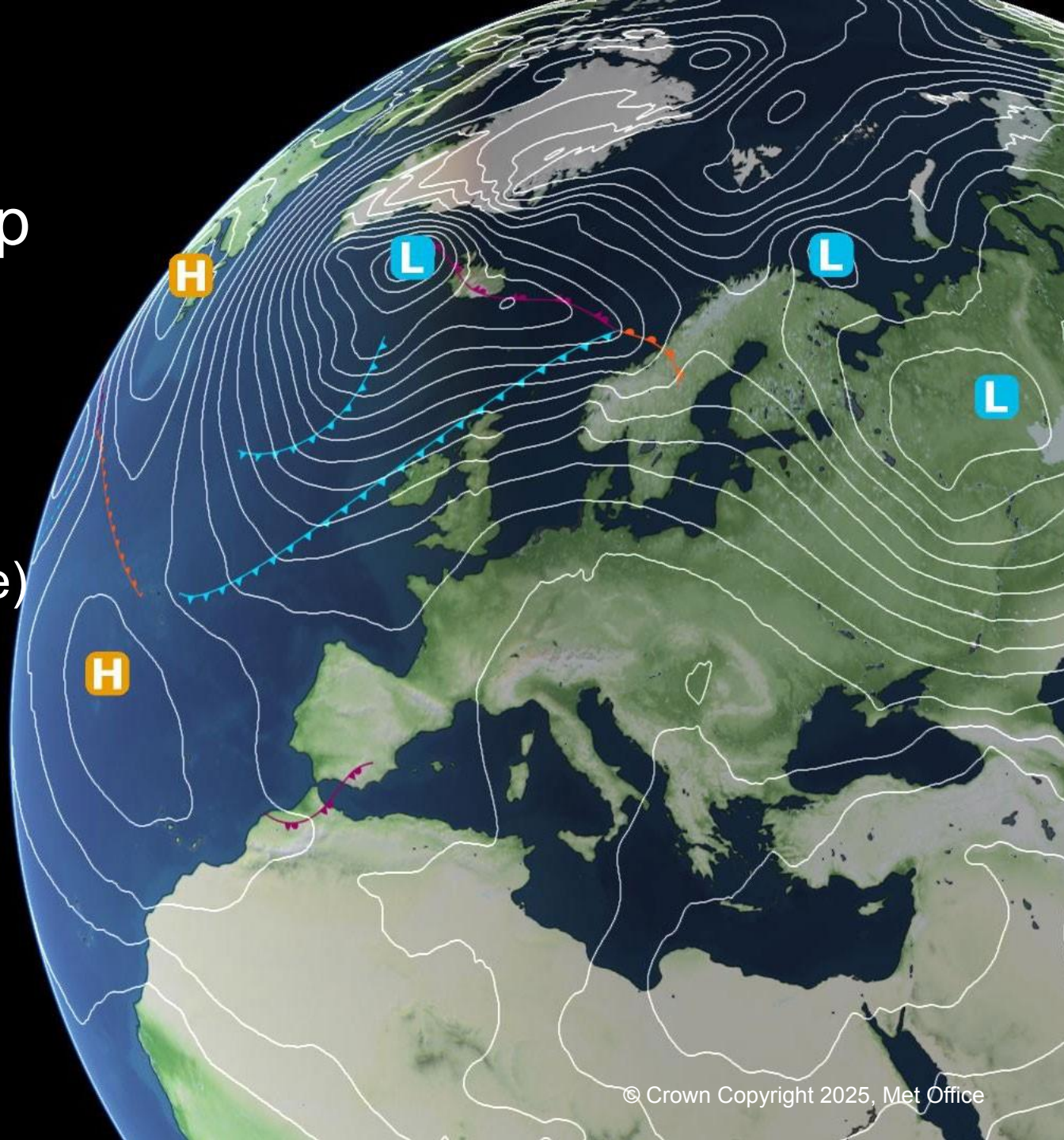
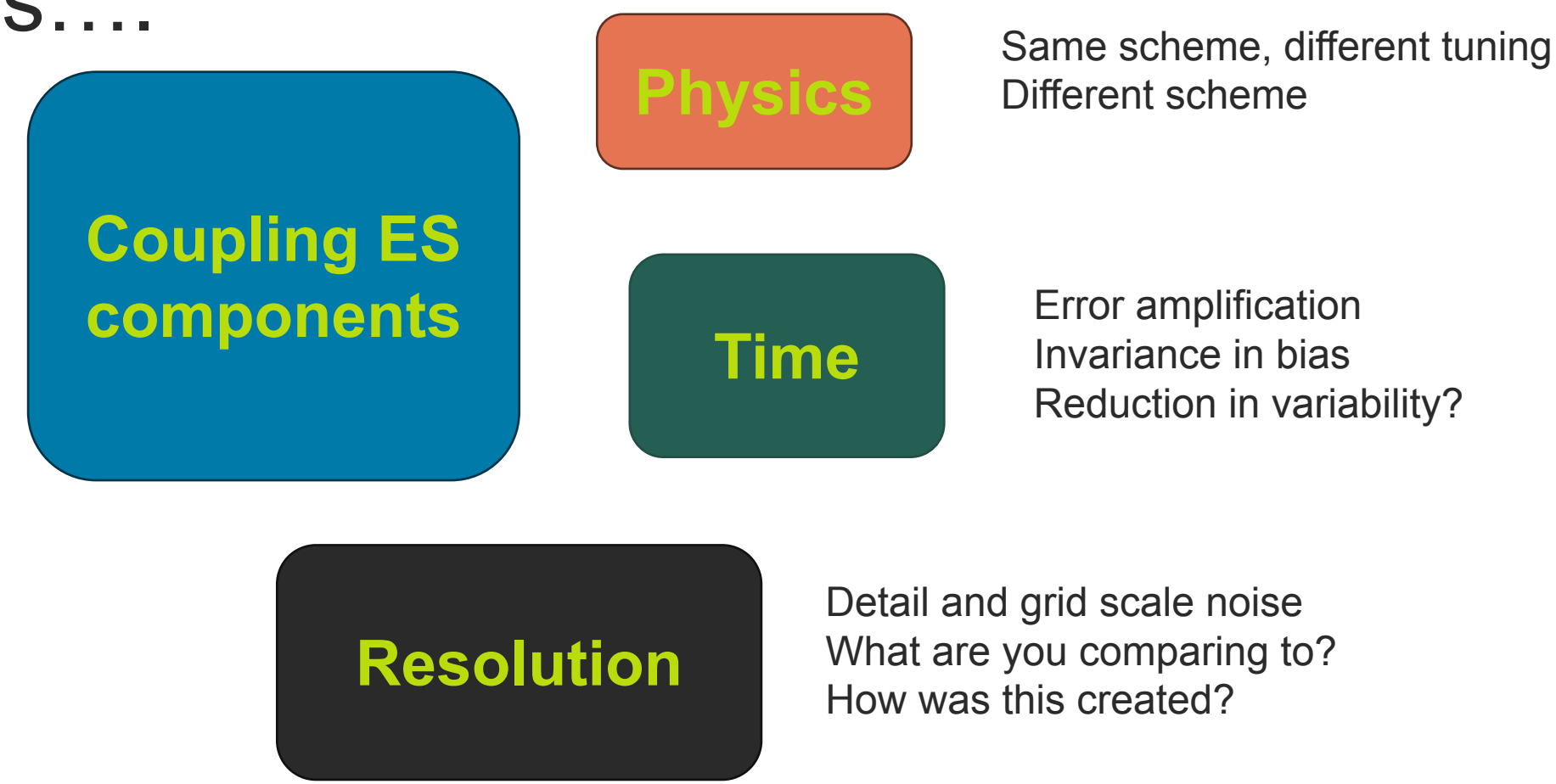
Can we use ensembles to help understand *(seamless)* model uncertainty?

Marion Mittermaier

(drawing on materials from many people)

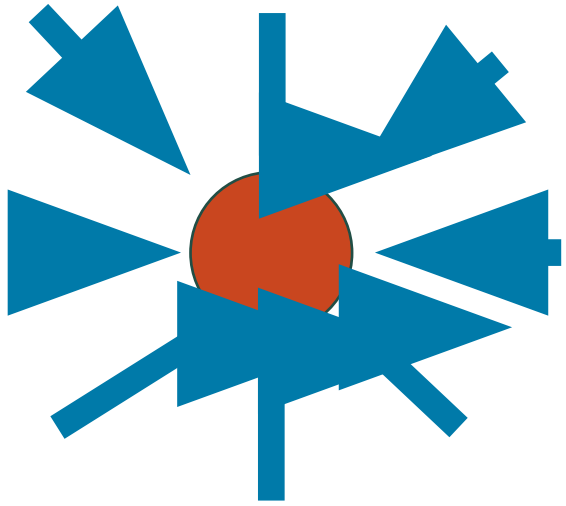


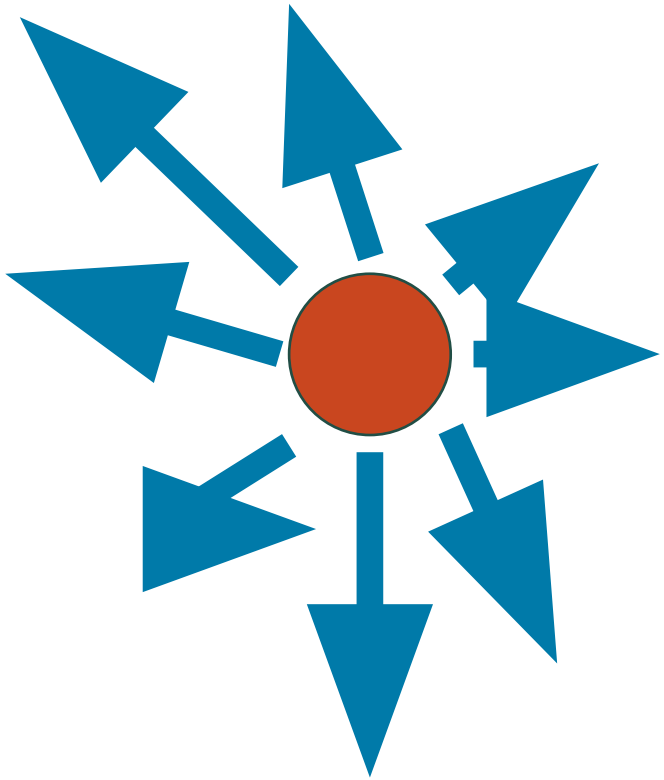
… and before we proceed any further, a word on seamlessness….



…. all these aspects can affect the bias and/or the variance

“ensemble” is an “overloaded class”





Perturbed parameter ensembles (PPE) constrain

NWP ensembles actively inflate or expand

Ensembles at different time scales are very different “beasts”

What are we running ensembles for?

Rare vs extreme vs high-impact

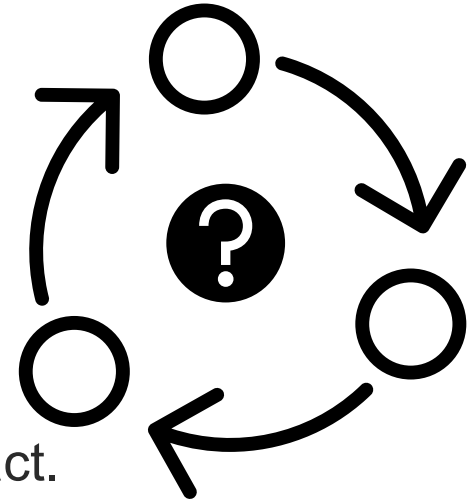
For longer time scales, trends and increasingly erratic behaviour are critical indicators for potentially more extreme scenarios.
At short-time scales we want to pinpoint the location and timing of impacts.

For NWP want maximum spread in super-quick time. This means "kicking" the model "hard".

For longer lead times we want the minimum spread which encapsulates a reliable signal of change.

Then the conundrum.

- Not all rare events are extreme.
- Not all extreme events are rare.
- Not all rare events are high-impact.
- Not all high-impact events are extreme….



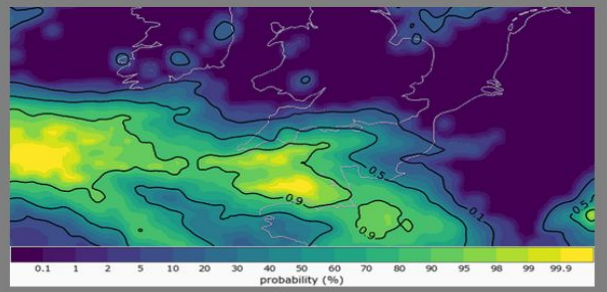
“Ensembles at the heart of everything we do”

Use cases for ensemble stats

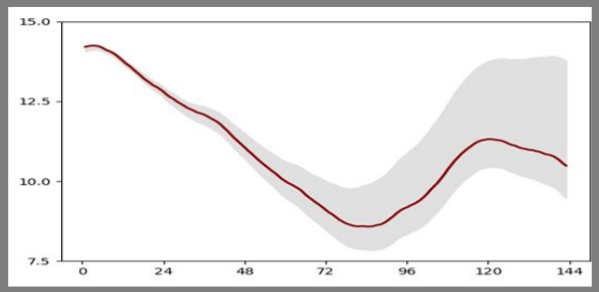
Best estimated value(s)

09:00	10:00	11:00	12:00	13:00
10%	10%	10%	40%	40%

Probability distributions

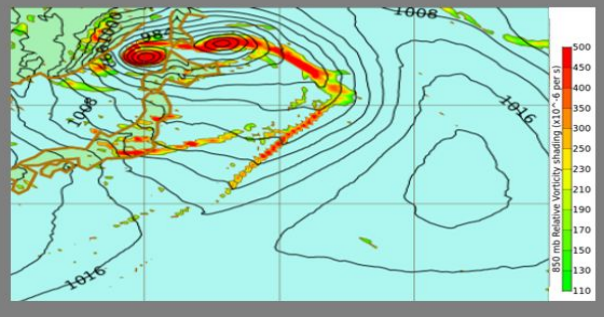


Range(s) of uncertainty

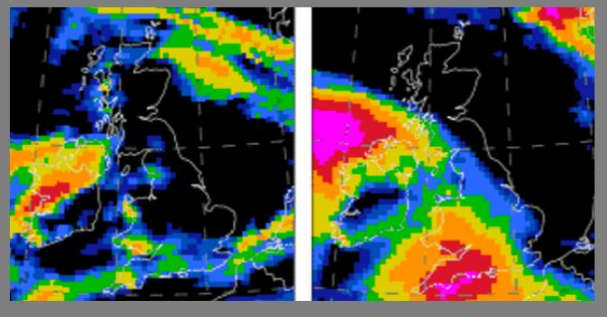


Use cases for one or more member

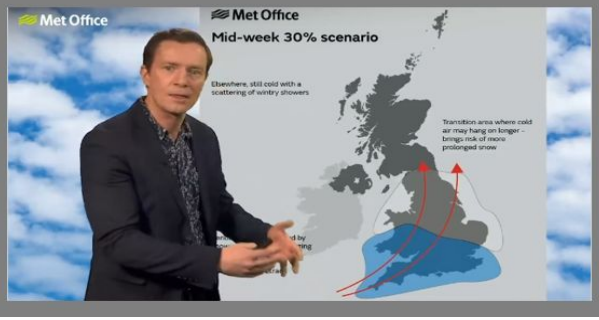
Representative member



Best/worst case scenario



Probable “storylines”



These are all NWP examples. Broadly speaking there are analogs for longer time horizons.

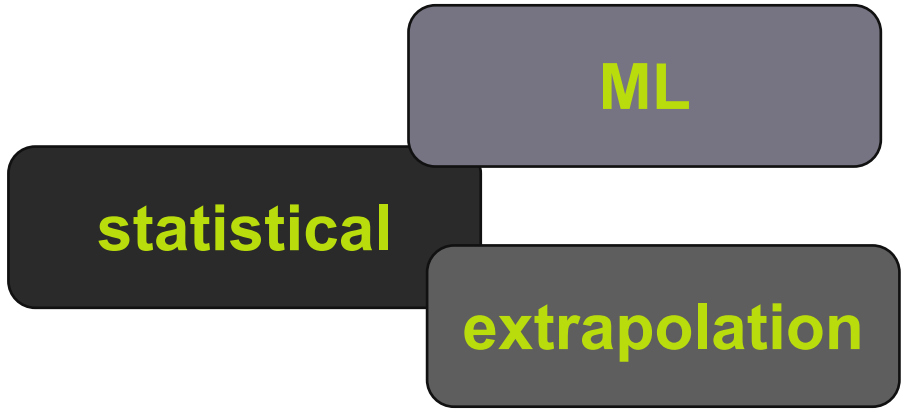
“model” has multiple meanings ….

Internal – sensitivity

physical

Different physics

Inter-model



Assumptions
Approximations
 All the outputs are *estimates*

Finally, “uncertainty” can be “woolly” too!

We are often “sloppy” with our language.
Uncertainty of what precisely?

Uncertainty ≠ bias

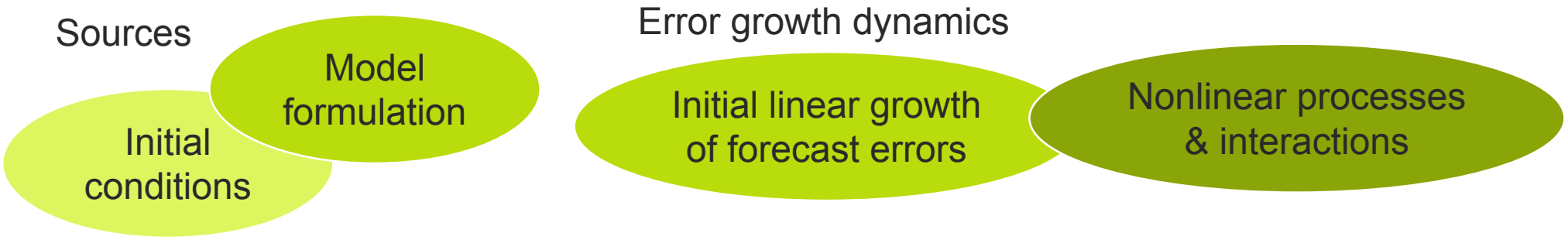
Bias = mean difference from an observed “state” (which may also be biased)

For ensembles we want to “equalise” the magnitude of the error and between-member variance.

So, what do we *mean* by **ensembles + model + uncertainty**?

Met Office

Ehrendorfer's (1997) *"Predicting the uncertainty of numerical weather forecasts: a review"*



Ensembles provide a practical way of predicting the time evolution of the forecast uncertainty.

Ensemble methods are considered the only feasible way to estimate the uncertainty in forecast states, given the considerable day-to-day variability in the forecast error.

Forecast uncertainty ≠ model uncertainty

Epistemic vs aleatoric uncertainty

Model (epistemic) uncertainty comes from a lack of knowledge or data.
Can be reduced with more data/knowledge.

High spread high epistemic uncertainty
Low spread model agrees and high confidence

Aleatoric uncertainty encompasses the noise in the data, e.g. observation sensor error.
Irreducible.

Affects quantification of spread.
*Makes an ensemble look (**more)** under-spread (than it actually is).*

From the user perspective there is a problem though…

High spread high epistemic uncertainty
Low spread model agrees and high confidence

Spread is desirable

Too much spread no guidance undesirable

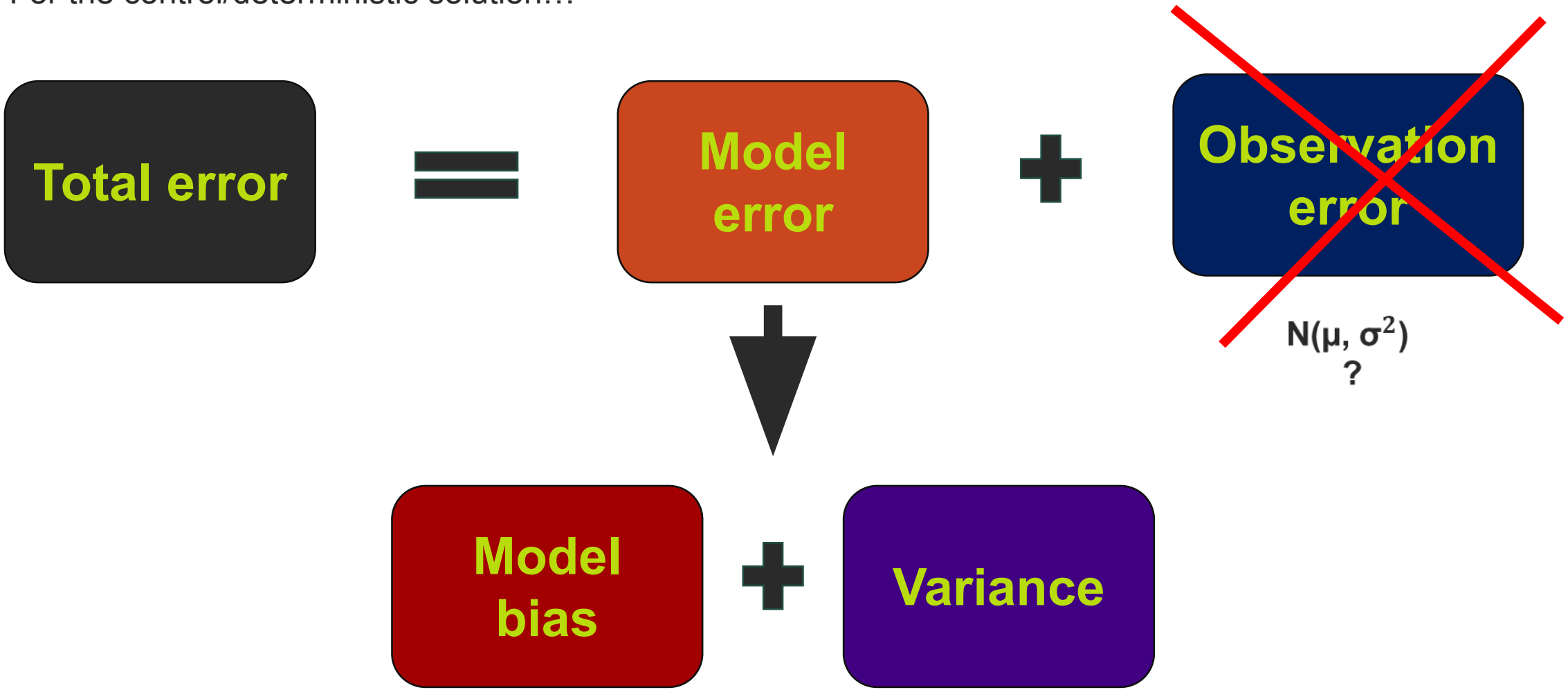
Ensemble members follow control undesirable

Sharp probabilities desirable

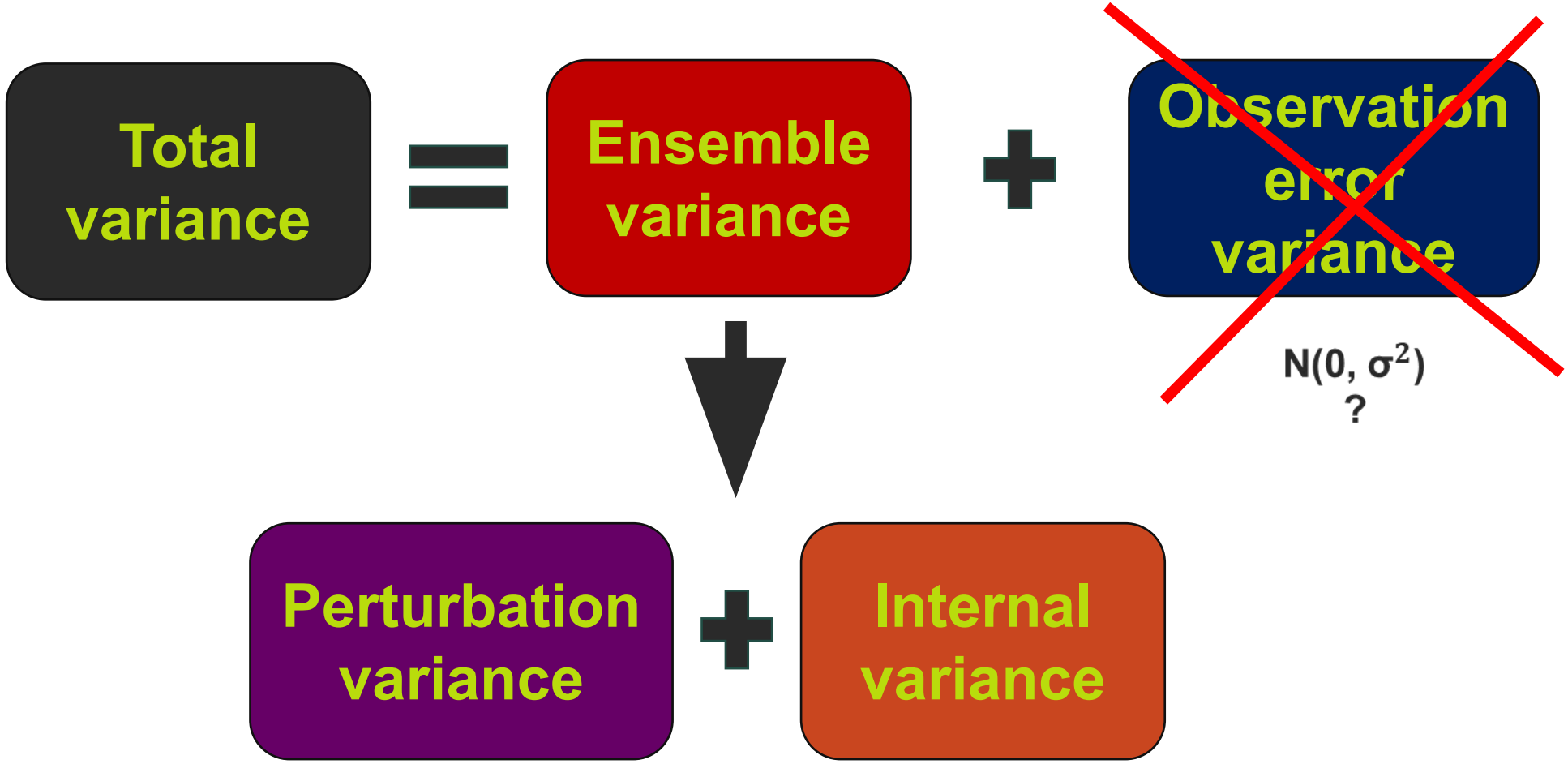
- All ensembles suffer from having finite (limited / few members)
- Ensemble size can be optimised (see Charlie's talk)

Total error hides true performance

For the control/deterministic solution…



For ensembles it gets even more convoluted



We typically get this from SPPT/SPP, SKEB, EnKF etc

What the model physics capable of producing on its own

How do you disentangle that?

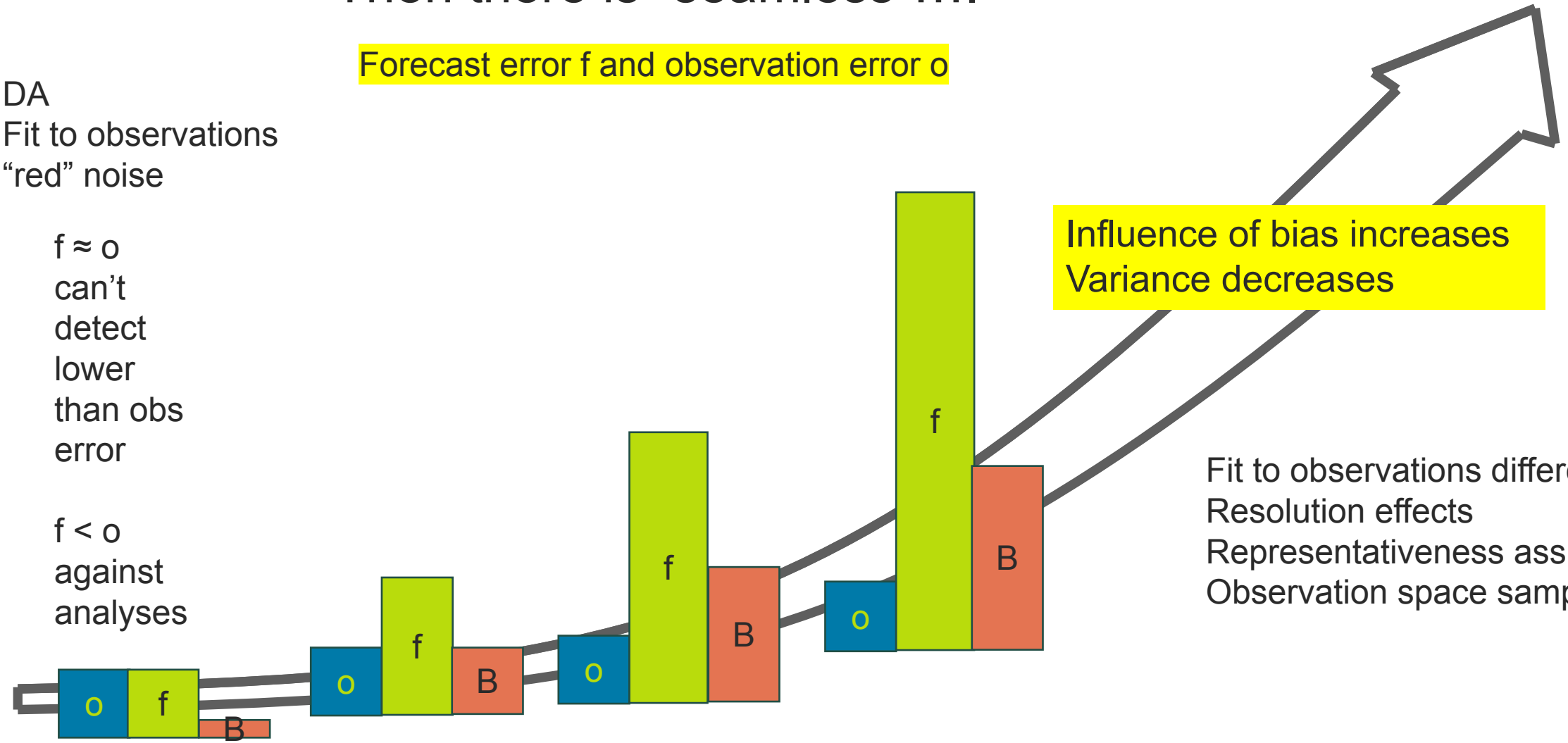
Then there is “seamless”….

Forecast error f and observation error o

DA
Fit to observations
“red” noise

f ≈ o
can’t
detect
lower
than obs
error

f < o
against
analyses

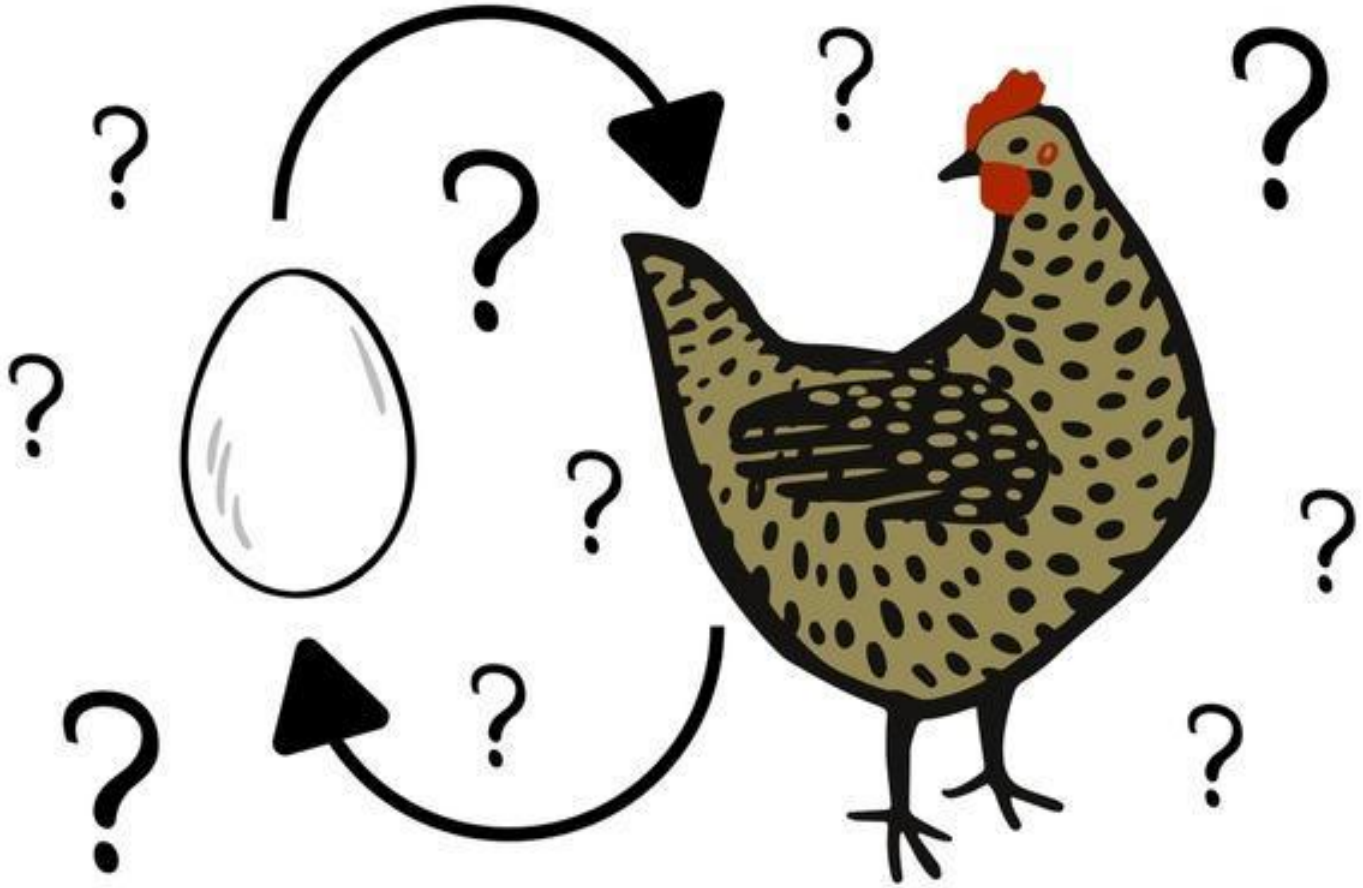


Influence of bias increases
Variance decreases

Fit to observations different.
Resolution effects
Representativeness assumptions
Observation space sampling

Nowcasting | 6-72h | 3-10 days | 10-30 days | 1-6 mo | Decadal | Climate

This still leaves one final question…



We have quite a few metrics, but they are bulk metrics.

Getting into the root causes of e.g. lack of spread and tuning sensitivity requires more than metrics.

It is a matter of experimental design.

There are (at face value) THREE parts to an egg, but many more parts to a chicken?!
What can we infer about the chicken from the egg?

Ensemble verification (at NWP timescales)

There are some subtle differences between ensemble and probabilistic forecast verification.

For ensembles, important concepts include:

- spread in ensemble member solutions
- error of the ensemble mean

We derive probabilities from ensembles. These may be raw (frequentist) or post-processed.

Note: Even the derivation of raw probabilities is a form post-processing!

Ensemble verification: a two-stage process

The underlying model configuration used to build the ensemble remains just as important to evaluate.

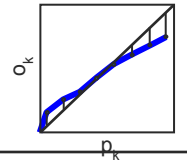
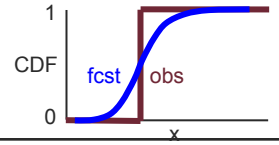
1. The **control** (which represents the core model) should be evaluated the same as a deterministic model to understand physical biases and systematic behaviour. This is the physical basis for the forecast and often represents our best estimate of the future weather. *[Not covered here.]*

2. The **impact of the perturbations used to create an ensemble forecast** and on forecast / ensemble performance is the second stage: it affects attributes such as the spread (variety/evolution of solutions), probability bias (reliability) and ability to discriminate between events and non-events probabilistically.

This second step is separate (and follows on from) gaining an understanding of the model itself.

Met Office

A side-by-side guide of metric analogs

Deterministic	Ensemble/Probabilistic	Visual aid
Mean bias	Reliability term of BS $\frac{1}{N}\sum_{k=1}^{K} n_k (p_k - \bar{o}_k)^2$	
RMS error	Brier score (square root) $\sqrt{BS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2}$	
Mean absolute error	CRPS $\int (P_{fcst}(x) - P_{obs}(x))^2 dx$	
Correlation	R^2 for logistic regression	

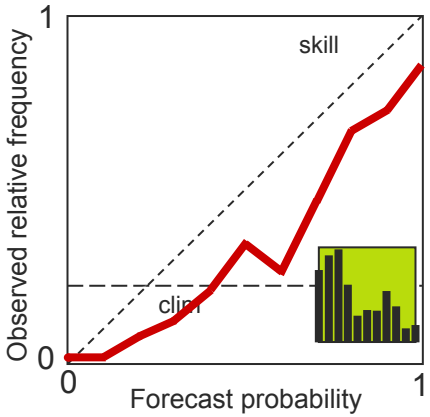
JWGFVR

A robust ensemble evaluation workflow

For probability forecasts

1. Reliability diagram

measures probability bias



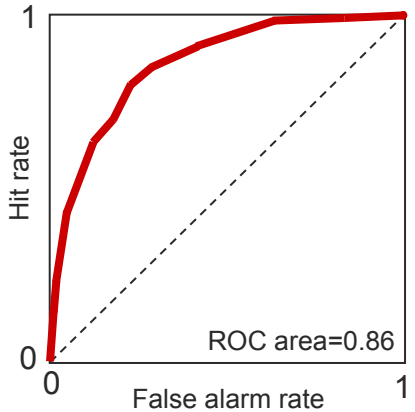
2. Brier score

measures probability error

```
Brier Score (BS) = 0.1445
Brier Skill Score = 0.1942
```

3. ROC

measures discrimination (potential skill)

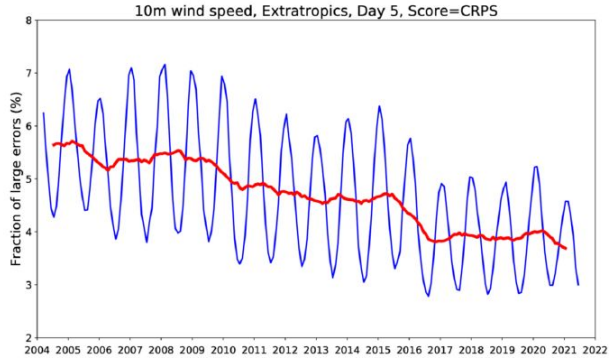


Fair scores

For ensembles you also need:

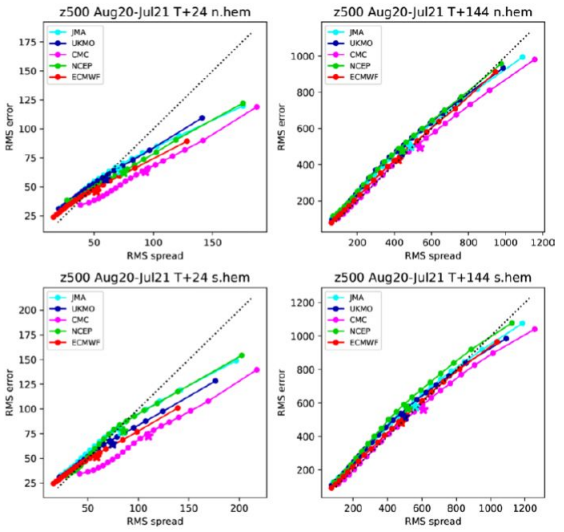
4. CRPS

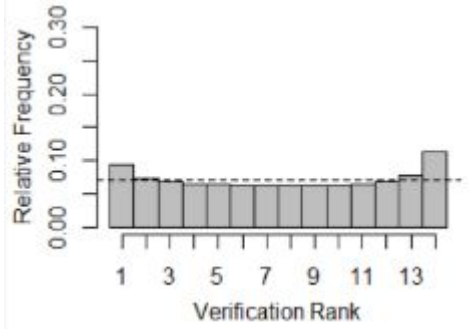
measures ensemble distribution



From ECMWF Tech Memo 884

5. Measures of spread-(skill)





From ECMWF Tech Memo 884

In summary, we have the tools…. Including spatial ones

Met Office

Ensembles and spatial methods

***Most* methods build on the important properties for ensemble metrics such as propriety.**

Spatial Probability Score (SPS) which is an integral of the BS along probability contours

eSAL an adaptation of SAL which utilises the CRPS

HiRA whilst one might think that an ensemble removes the need for spatial neighbourhoods, it is clear that for higher resolutions a small neighbourhood remains beneficial (though larger neighbourhoods are not). Best to use with a fair score.

MODE measure spread-error in object/feature attributes not proper

Neighbourhood-based scores based on CRPS (Stein and Stoop)

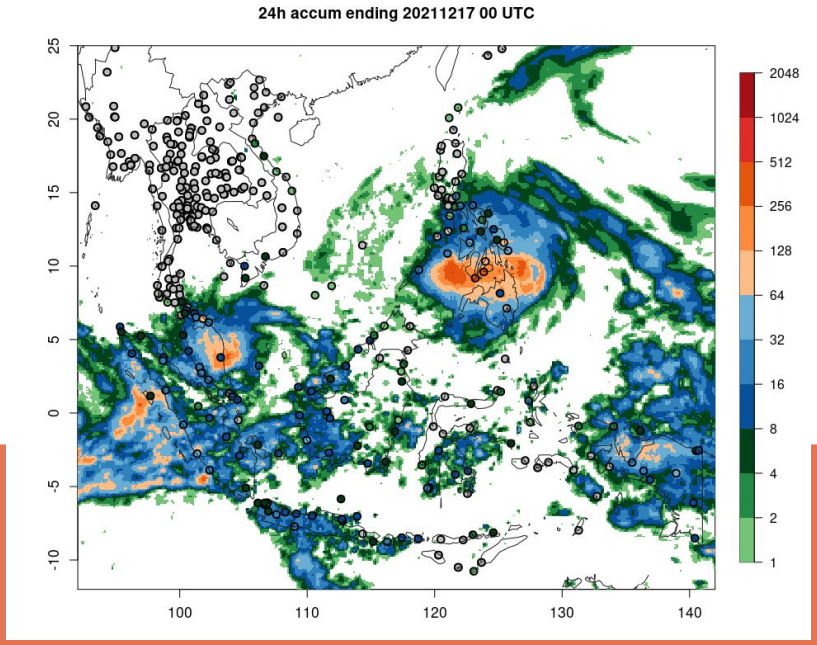
Not proper:

FSS boils down to an evaluation of the ensemble mean; spread skill with dFSS-eFSS

Ensemble-specific tensions

- ***Biases*** in the physical model can be very detrimental for an ensemble.
- If they are diagnosed as being systematic, they **can be fixed with post-processing of the ensemble *members***, *before* any probabilities are calculated.
- However, probabilities can *also* be post-processed! This is *not* the same thing.
- Tension between *spread* and *sharpness (we want both, this I would call the equivalent of having cake and eating it!)*

Observation uncertainty



For the evaluation of ensembles, observation uncertainty is particularly problematic.

All observations have errors.

Comparing different observations of the same quantity at the same time often displays differences. So, which observation is right? Can we say?! (no)

We want to know whether the forecast is right, but how can we do that when we don't have 100% confidence in the observed state?

This is bad for a yes/no deterministic forecast, but worse for ensembles.

Ensembles are used to get a measure of forecast uncertainty.

What is forecast uncertainty and what is observation uncertainty? Or internal model uncertainty?

Affects all aspects of ensemble performance assessment, especially our ability to assess reliability and spread-skill, and most acute at shorter lead times.

Effect of observation uncertainty

Observation errors and uncertainty add uncertainty to the verification results

- True forecast skill is unknown
  - → An imperfect model / ensemble may score better!
- Extra dispersion of observation PDF

Typical effect on verification results:

- RMSE – *overestimated*
- **Spread – more observed outliers make ensemble look under-dispersed**
- Reliability – poorer
- Resolution – greater in BS decomposition, but ROC area poorer
- CRPS – poorer mean values

Can we remove the effects of observation error?

- Add observation error to ensembles before verifying (Candille and Talegrand)
- More samples helps with reliability estimates
- Error modelling – study effects of applied observation errors
- Need "gold standard" to measure actual observation errors

Accounting for Representativeness in the Verification of Ensemble Forecasts

Zied BEN BOUALLEGUE

Motivation

Ensemble forecasts provide information about **forecast uncertainty.**

When verifying ensemble forecasts**, not** accounting for **observation uncertainty** leads to:

- encouraging forecasts of erroneous observations (instead of the truth)
- inaccurate skill estimation
- misleading comparison between forecasting systems

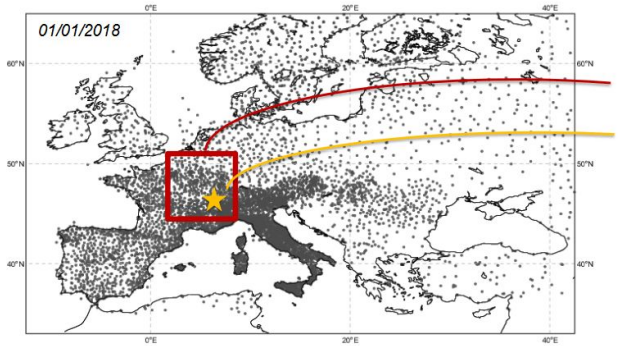
Focus on **"representativeness"**

- **definition**: mismatch between a quantity measured at two different scales
- **assumption**: representativeness is the principal source of observation uncertainty
- **application**: global ensemble forecast verification of surface variables

Observation uncertainty characterization

based on **observations only**, a network of **high-density observations** (over Europe)

➢ is a single location measurement ★ representative of an average ☐ over larger areas?



box-average y_A

point-observation y_B

$P(Y_B | Y_A)$?

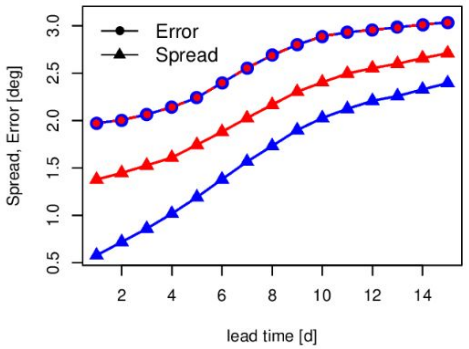
ECMWF

Spread/skill relationship - Forecast deficiencies

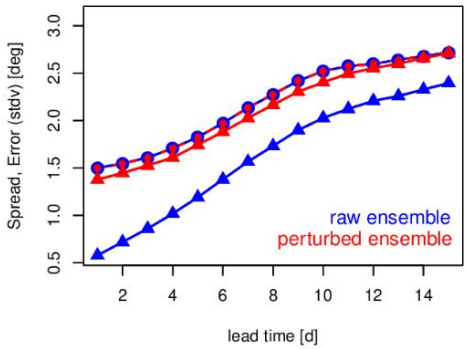
Good spread-skill relationship when comparing stdv of the error and spread of the perturbed ensemble

- perturbed ensemble accounts for representativeness error in the observation, **not in the forecast**

No bias correction



With "bias correction"



Why it matters

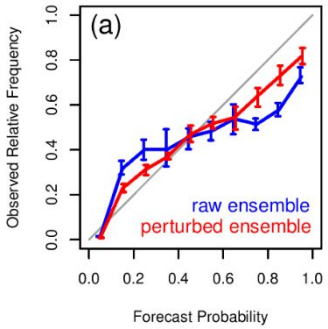
Impact on scores

Large impact on **reliability/sharpness** attributes

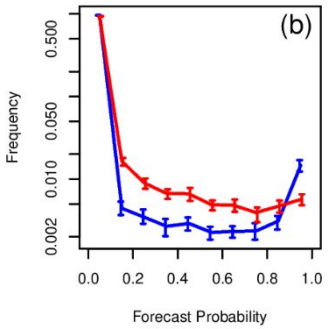
❑ Summer 2018 - Europe - 00UTC - day 5 - threshold: 25 deg.

➢ **more reliable** probabilistic forecasts with observation uncertainty

Reliability



Sharpness

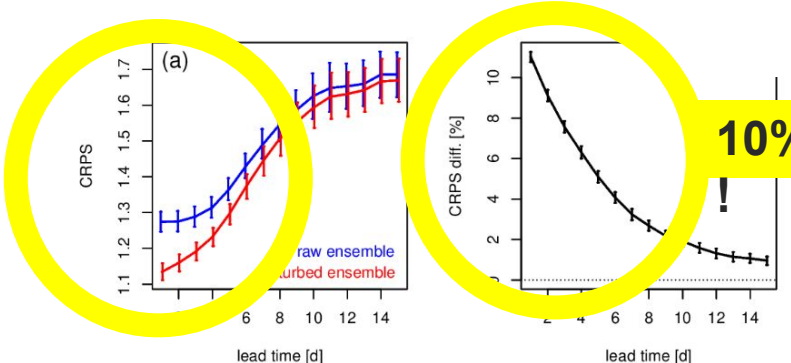


Impact on scores

Comparison of the **raw ensemble** vs the **perturbed ensemble**

❑ Summer 2018 - Europe - 00UTC

➢ **smaller error** with observation uncertainty



10%!!!

JWGFVR

What model development groups said

A set of questions were sent far and wide to Met Services and model development groups to ask them some fundamental questions.

Responses so far from DWD, NCAR (MPAS), Met Office ….

Spoiler: There is a pattern emerging!

Met Office

1. Do you have a “recipe” that you follow for the development process and what evaluation supports this.

… the amount of spread crucially depends on model error, **an essential contribution to ensemble development is improving the deterministic model and keeping the ensemble in line**. To simulate the remaining errors, we use stochastic SST and parameter perturbations. The latter in the sense of a **multi-model approach to ensemble forecasting.**

SSA develop our ensemble system **by trying to estimate model error during the DA 6-hour window and trying to get the spread to match** that through various concoctions like inflation (RTPP, RTPS). In R2O we separately look at model forecast error and spread (taking the initial perturbations developed by SSA’s VAR scheme as given), **and try to maximise spread from stochastic physics, additive inflation, SST perts, soil-moisture and other land-surface perturbations** etc. This is usually from RMSE vs Spread line plots or **more recently looking at maps of spread and error.**

No recipe. We kind of do what we want. We just try things out, verify the forecasts, and try to fix problems**. It’s a fairly iterative process.**

Met Office

2. How much emphasis is put on the deterministic model (control) performance (i.e. the physics) before turning it into an ensemble?

Very much. We are **trying to keep biases of the ensemble mean small and as close as possible to the high resolution deterministic model**. If the ICON model changes, the score cards of both operational systems are checked and we pay specific attention to changes in ensemble spread in the verification against observations as well as analysis fields.

In a GA/GC release **we focus almost exclusively on the deterministic performance and then check that the ensemble is following suit afterwards.** To date our concerns has been model stability in the ensemble and to not lose more spread.

We frequently **initialize limited-area ensembles from our own cycling DA system** (per previous paper). And as part of that, **we carefully look at the prior (before DA) statistics in observation space**. This allows us to assess model biases and characteristics of the spread.

Met Office

3. Do you do anything special to assess the interaction between the DA and the model and the ensemble (especially if the DA is involved in helping create the perturbations)?

ICON **adaptive parameter tuning related to DA increments**

We take our **global LETKF analysis ensemble to initialize the ICON-EPS forecasts**. The analysis ensemble spread **includes co-variance inflation to stabilize the DA cycle and is tuned to give sufficient spread in the medium range**. Even though we use stochastic SST-perturbations at initial time and during the forecast, we still report a spread/skill ratio of 1.7 at initial time to WMO ensemble verification.

Singular Vectors have been further developed in the past years

This is done when SSA developed the ETKF and then later the En-4dEnVar scheme.

We **mostly use ensembles to produce and verify probabilistic forecasts. This allows us to understand model uncertainty and spread**. We don't look at individual member solutions very much, other than perhaps to look at things like convective modes and structures.

Met Office

4. What do you understand by the following: “using ensembles to understand model uncertainty and spread”? Do you use ensembles this way, and how? If not, can you think of ways on how you could use ensembles to do this?

Yes, we use ensembles this way. **We are running ensemble experiments in our model developing and verification suites to understand how model changes impact forecast uncertainty and ensemble spread.**

It is meant to be an application of ensemble output to help develop new physics configurations. We don’t do this at the Met Office in any meaningful way in NWP. They do something like this in PPE (perturbed parameter elicitation) work for decadal predictions.

Met Office

5. Does any of this have any relevance in an ML world?

The **operational implementation of the Singular Vectors will be used together with small scale noise and perturbations from DA to set up a reliable AICON-EPS**. This is a contribution to WP2 "ML Ensembles" in the MLPP.

SVs are specifically helpful in understanding ML systems and contribute to XAI by systematically exploring the "dynamic" sensitivities of MLWP models, which can be compared to those of NWP.

I think it depends. If considering AI NWP models, I'm not sure if it's relevant. Maybe. **But if using ML as a postprocessing tool to effectively correct raw forecasts, then I do think aspects of ensemble design are important to give the ML system the best possible states.**

So, what about seamlessness then?

- *We currently do not use ensembles to evaluate uncertainty in a seamless sense.*
- **Deterministic evaluation remains the cornerstone for the core model across all time scales.**
- **Things are looking more promising** between global and regional evaluation, though this is still predominantly deterministic, thanks to the move to km-scale global modelling.
- On a “macro” level it would appear as though there is **very little difference in how to “do” ensemble verification**, including estimating model uncertainty **across NWP and S2S time scales**.

Beyond that? It is a very different proposition.
Can we become seamless?

Ensemble evaluation specifics

- Ensemble and probabilistic forecast verification is far **more nuanced** than deterministic forecast verification.
- **Robustness of metrics is fundamentally important**: propriety, fair scores etc.
- **All time scales can / would benefit from more spatial assessments**: e.g., spread/skill and greater focus on event-based evaluation.
- As always, **never rely on one metric or one method**.
- **We should not be alarmed if different methods (or observation sources/types) give different results**. Using the methods must come with an understanding that each method measures slightly different attributes. The apparent contradictions may therefore not actually be contradictory but pointing to different aspects of model behaviour.
- **For NWP time scales, observation uncertainty can play a large part** and must be accounted for to interpret ensemble/probabilistic statistics safely (unambiguously).
- **Observation uncertainty is still present at longer lead times** but will manifest itself differently. Whilst the mean behaviour is probably captured fairly well, the **variability may not be**. Resolution dependence. Cannot fabricate detail in observation data sets that isn't there. Unknown unknowns.

What steps can we take towards using *ensembles* to *seamlessly understand model uncertainty*?

- We **need to tighten up our language** on what we mean by model uncertainty. It can be ambiguous, particularly across NWP and climate.
- Much of the **methodology** to assess the impact on spread and uncertainty is there.
- In terms of the classic spread-skill line graph, it is unclear to me how you can differentiate sources of uncertainty without more targeted sensitivity experiments to test the sensitivity to individual parameters.
- Fundamental to doing this is the **right experimental design and model hierarchy**. PPE seems to provide a good template for longer time scales. Can we make this, or something similar work for shorter ones to give more targeted output. Is it affordable?
- Consider how the experimental designs need to be adapted for ML or hybrid models.


Met Office
Thanks for listening!
www.metoffice.gov.uk
© Crown Copyright 2024, Met Office