The difficulty of verifying small improvements in forecast quality

Alan Geer

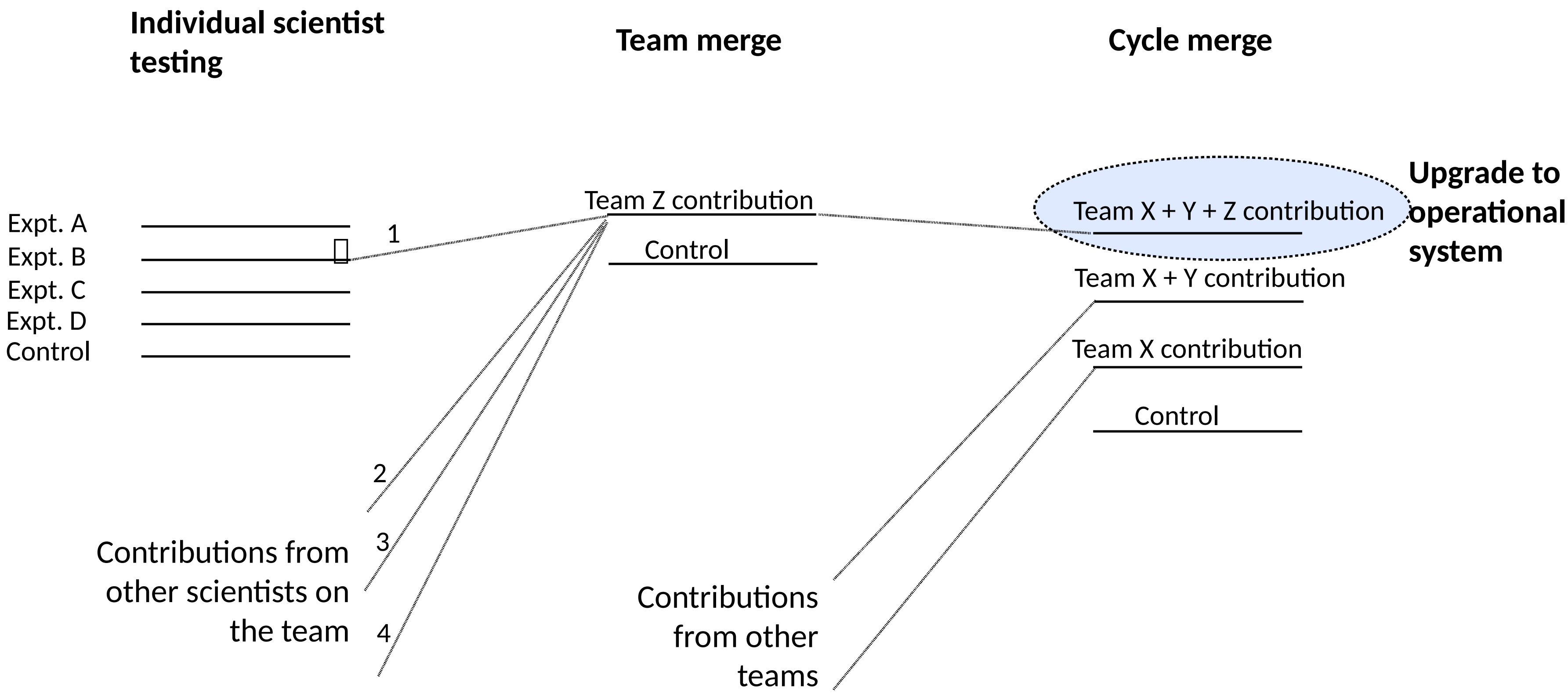
Satellite microwave assimilation team, Research Department, ECMWF
(Day job: all-sky assimilation)

Thanks to: Mike Fisher, Michael Rennie, Martin Janousek, Elias Holm, Stephen English, Erland Kallen, Tomas Wilhelmsson and Deborah Salmond

The viewpoint from an NWP research department

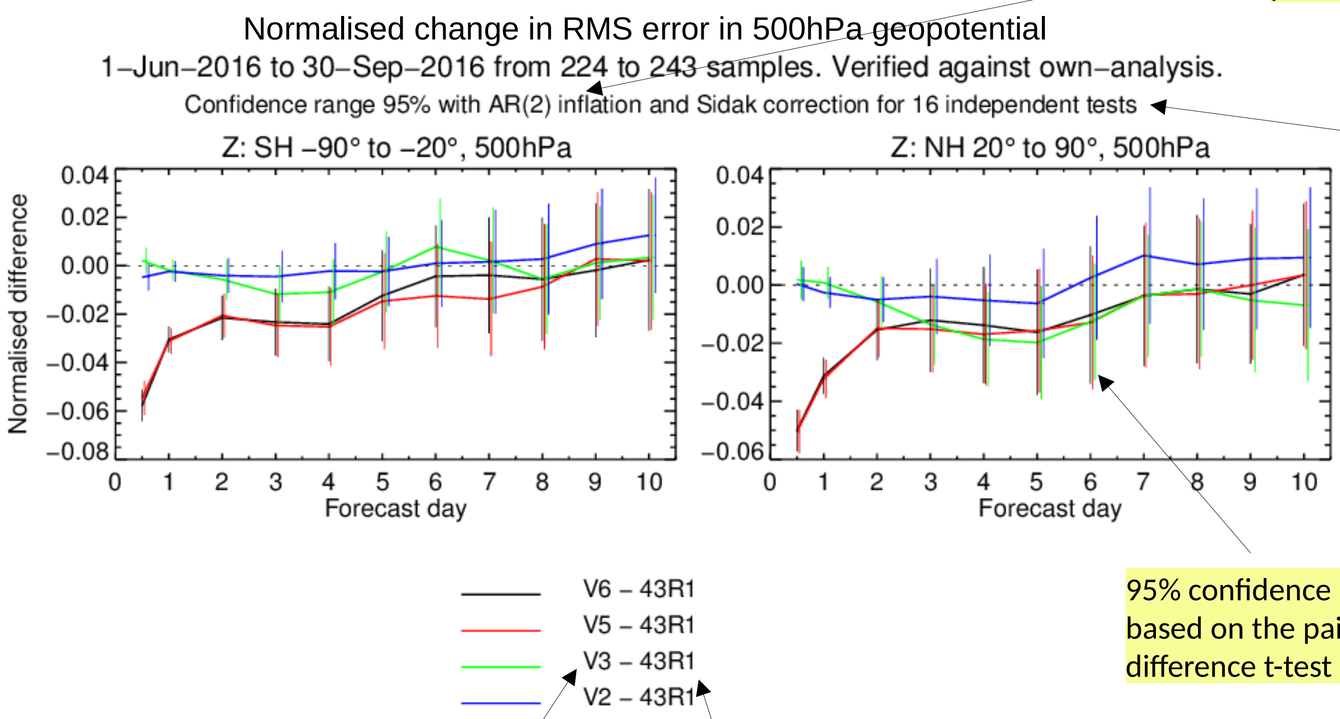
- Not:
  - What is the skill of a forecast?
  - Is one NWP centre's forecast better than another?
- But this:
  - Is one experiment better than another?
  - Is the new cycle (upgrade) better than current operations?
- Philosophy:
  - Lots of small improvements add up to generate better forecasts.

Research to operations



“Iver”: an R&D-focused verification tool

Confidence interval inflation to account for time-correlations of paired-diffences in forecast errors



Correction for multiplicity:

- 4 separate experiments
- 2 hemispheres
- roughly 2 independent scores across days 3-10

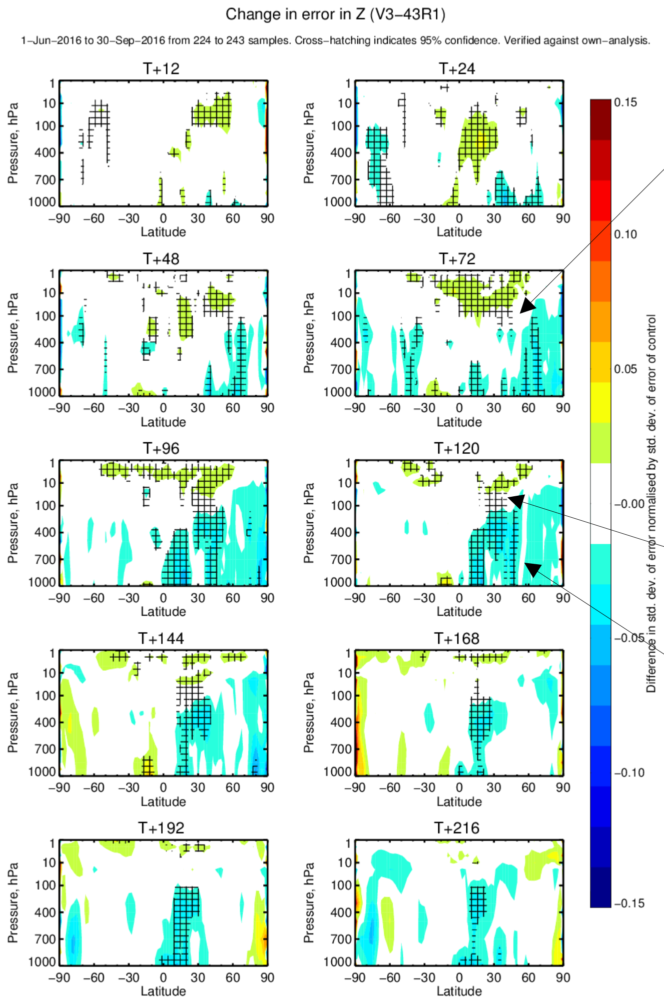
95% confidence intervals, based on the paired-difference t-test

Experiments progressively adding different components for the cycle upgrade

Control: current operational system (“Cycle 43R1”)

Latitude-pressure verification

Normalised change in std. dev. of error in Z (experiment - control)



A typical dilemma in NWP development:

- Should we accept a degradation in stratospheric scores to improve tropospheric midlatitude scores?
- Do we even believe the scores are meaningful?

Cross-hatching: significant at 95% using t-test with Šidák correction assuming one panel contains 20 independent tests

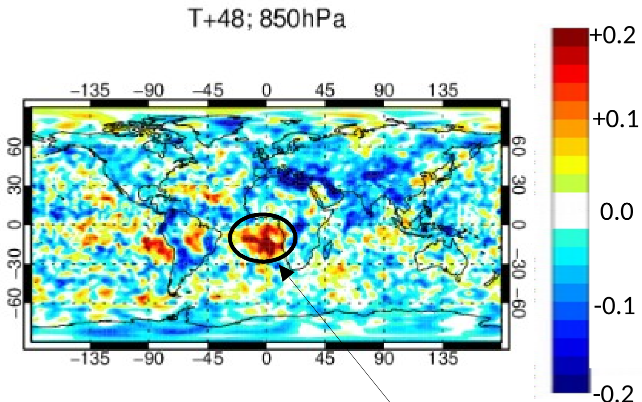
Blue = reduction in error = experiment better than control

ECMWF

Latitude-longitude verification

Because many improvements (and degradations) are local

Are these patterns statistically significant?
- requires multiplicity correction: work in progress



Normalised change in RMS T error at 850hPa

But are these patterns useful despite the lack of significance testing?

- Yes, this turned out to be a problem associated with a new aerosol climatology that put too much optical depth over the Gulf of Guinea
  - Too much optical depth = too much IR radiative heating at low levels = local temperatures too warm

Statistical problems in NWP research & development

- The issues:
  - Every cycle upgrade generates hundreds of experiments
  - NWP systems are already VERY good: experiments usually test only minor modifications, with small expected benefits to forecast scores
  - Much of what we do is (in the software sense) **regression testing:**
    - We are checking for unexpected changes or interactions (bugs) anywhere in the atmosphere, at any scale
    - Verification tools will generate 10,000+ plots, and each of those plots themselves may contain multiple statistical tests

- Accurate hypothesis testing (significance testing) is critical:
  - Type I error = rejection of null hypothesis when it is true = **false positive**. Can be more frequent than expected due to:
    - Multiple testing (multiplicity) 1
    - Temporal correlation of forecast error 2
  - Type II error = failure to reject null hypothesis when it is false
    - Changes in forecast error are small; many samples required to gain significance 3

- Are our chosen scores meaningful and useful? 4

1. Multiple comparisons (multiplicity)

- 95% confidence = 0.95 probability of NOT making a type I error
- What if we make 4 statistical tests at 95% confidence?
  - Probability of not making a type I error in any of the four tests is:

    $$0.95 \times 0.95 \times 0.95 \times 0.95 = 0.81$$

  - We have gone from 95% confidence to 81% confidence.
  - There is now a 1 in 5 chance of at least one test falsely rejecting the null hypothesis (i.e. falsely showing "significant" results)
- Šidák correction:
  - $P_{TEST} = (P_{FAMILY})^{(1/n)}$
  - If we want a family-wide p-value of 0.95, then each of the four tests should be performed at 0.987

Shouldn’t *n* be very large?

- If we generate 10,000+ plots, why isn’t *n*>10,000?
- Because many of the forecast scores we examine are NOT independent

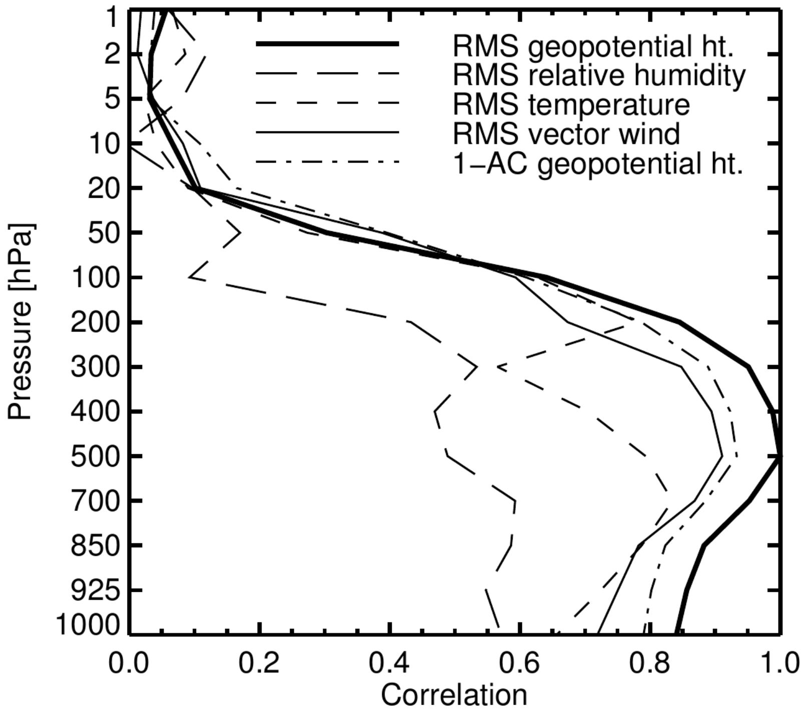
Testing the statistical significance testing

Geer (2016, Tellus): Significance of changes in forecast scores

- Three experiments with the full ECMWF NWP system, each run over 2.5 years:
  - **Control**
  - **AMSU-A denial:** Remove one AMSU-A (an important source of temperature information) from the observing system
  - **Chaos:** Change a technical aspect of the system (number of processing elements) that causes initially tiny numerical difference in the results, which quickly grow.
    - A representation of the null hypothesis: no scientific change

ECMWF

Correlation of paired differences in other scores with paired differences in day-5 Z RMSE scores

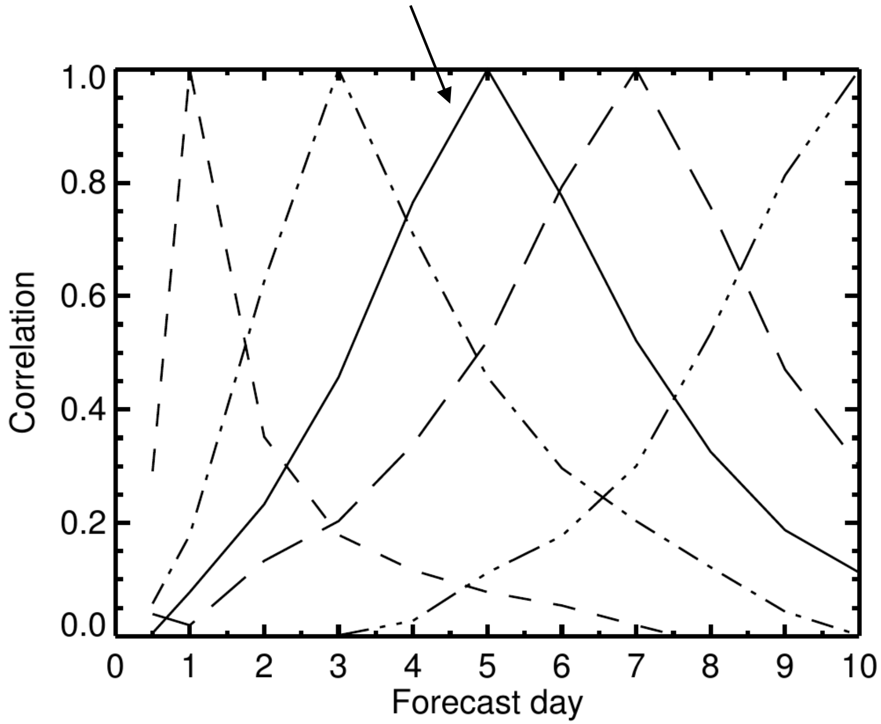


- All the dynamical scores are fairly correlated over the troposphere, and with one another

→ Z500 RMSE is sufficient to verify tropospheric synoptic forecasts in the medium range

But the stratospheric scores, and relative humidity, appear more independent

Correlation of paired differences in scores at other time ranges with paired differences in day-5 Z RMSE scores

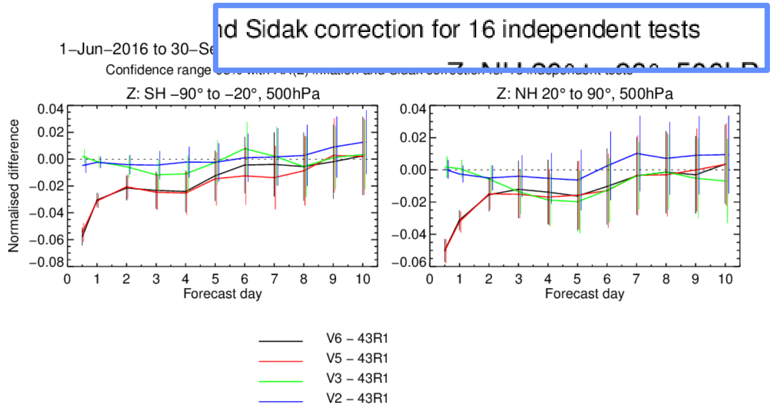


- Scores are correlated over a few days through the time range

→Day 5 RMSE Z is sufficient to verify the quality of (roughly) the day 4 to day 6 forecasts

What is a reasonable *n*?

- For the regional scores, *n* is the product of:
  - Number of experiments
  - Medium-range and long-range
  - Two hemispheres

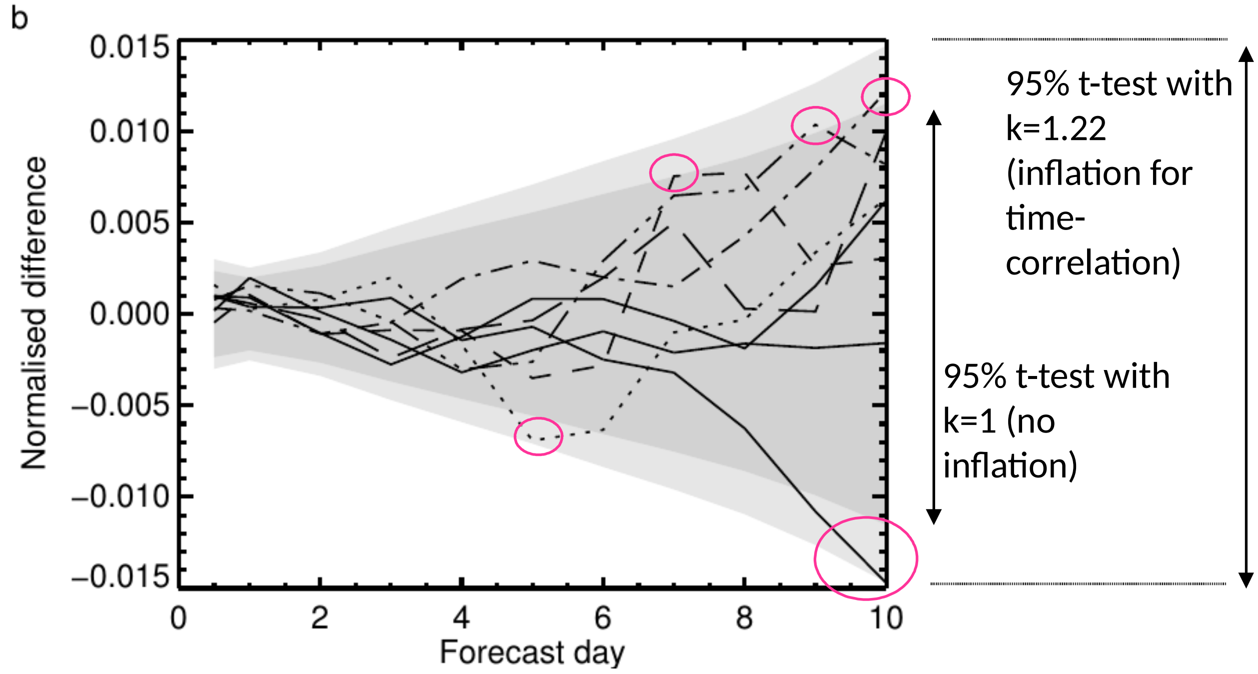


- But why not also count the stratosphere, tropics, lat-lon verification?
  - For the moment, *n* is computed independently for each style of plot

2. Type I error (false rejection of the null hypothesis) due to time-correlation of forecast errors

The chaos experiment should generate false positives at the chosen p-value (e.g. 0.95). Instead, naive testing generates false positives far more frequently.

Chaos – control, computed on 8 chunks of 230 forecasts



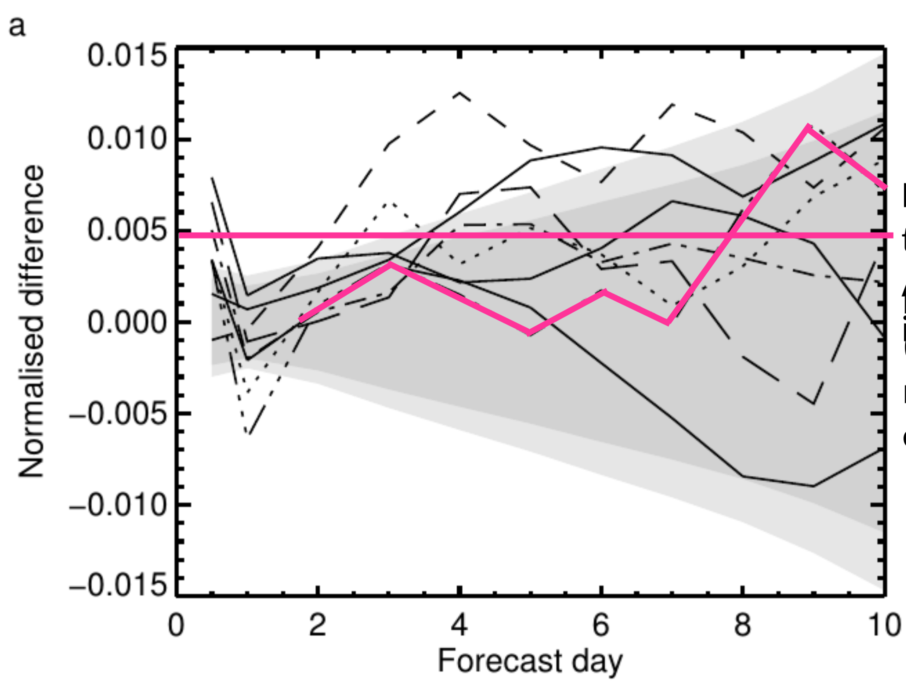
95% t-test with k=1.22 (inflation for time-correlation)

95% t-test with k=1 (no inflation)

3. Type II error: failure to reject the null hypothesis

The AMSU-A denial experiment should degrade forecast scores. AMSU-A is a very important source of data, known to provide benefit to forecasts

AMSU-A denial – control, computed on 8 chunks of 230 forecasts

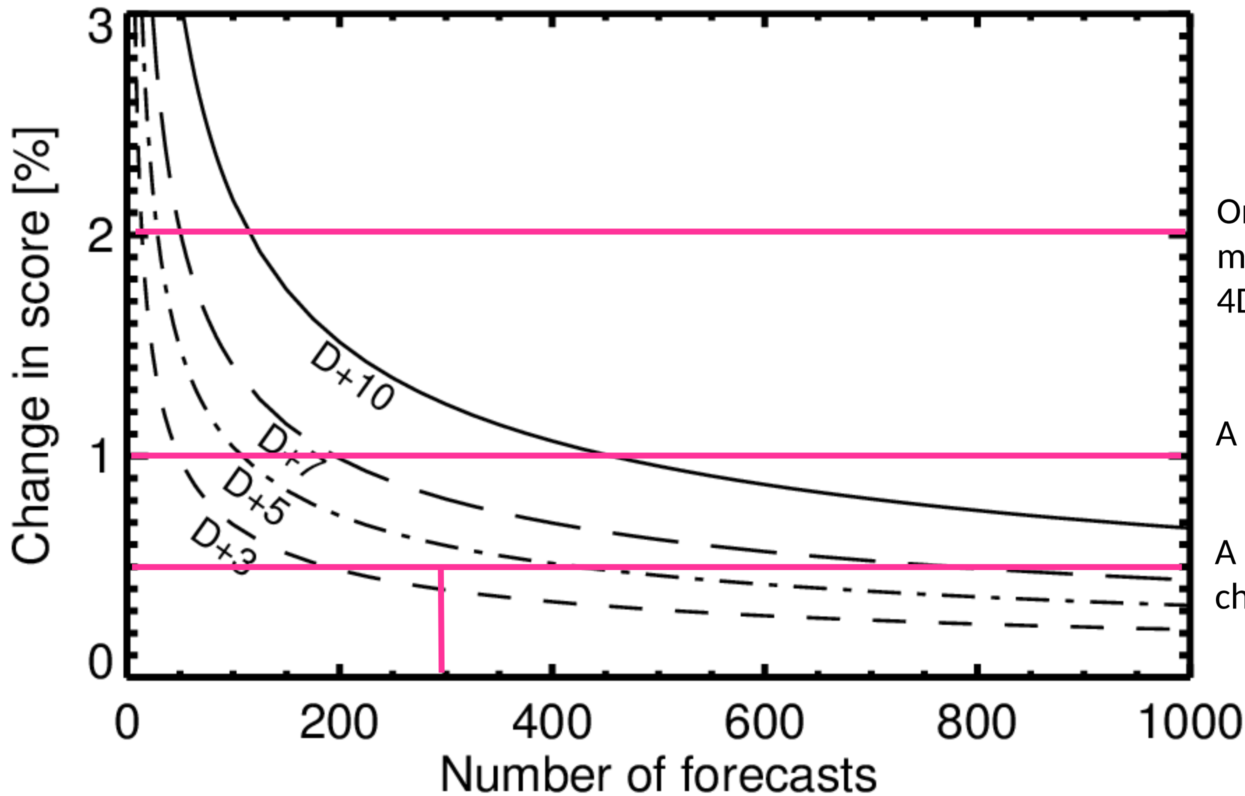


Based on 2.5 years testing, we know the AMSU-A denial impact is this

But on 230 forecasts (about 4 months) we might get this: Type II error

Fighting type II error: How many forecasts are required to get significance?

1 independent test (e.g. we have one experiment and all we care about is NH day 5 RMSE)



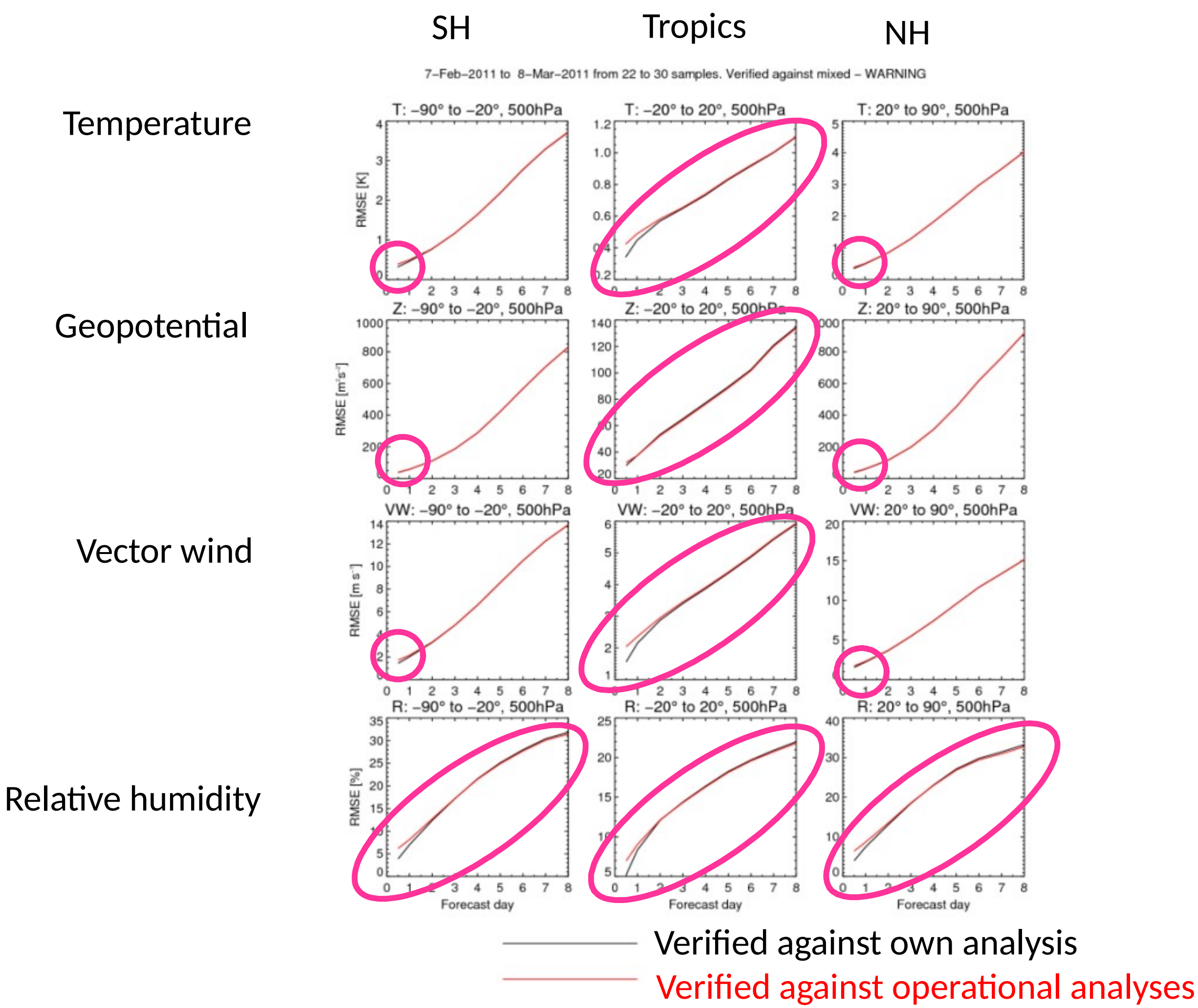
Once in a while (e.g. moving from 3D-Var to 4D-Var)

A typical cycle upgrade?

A typical individual change, e.g. one AMSU-A

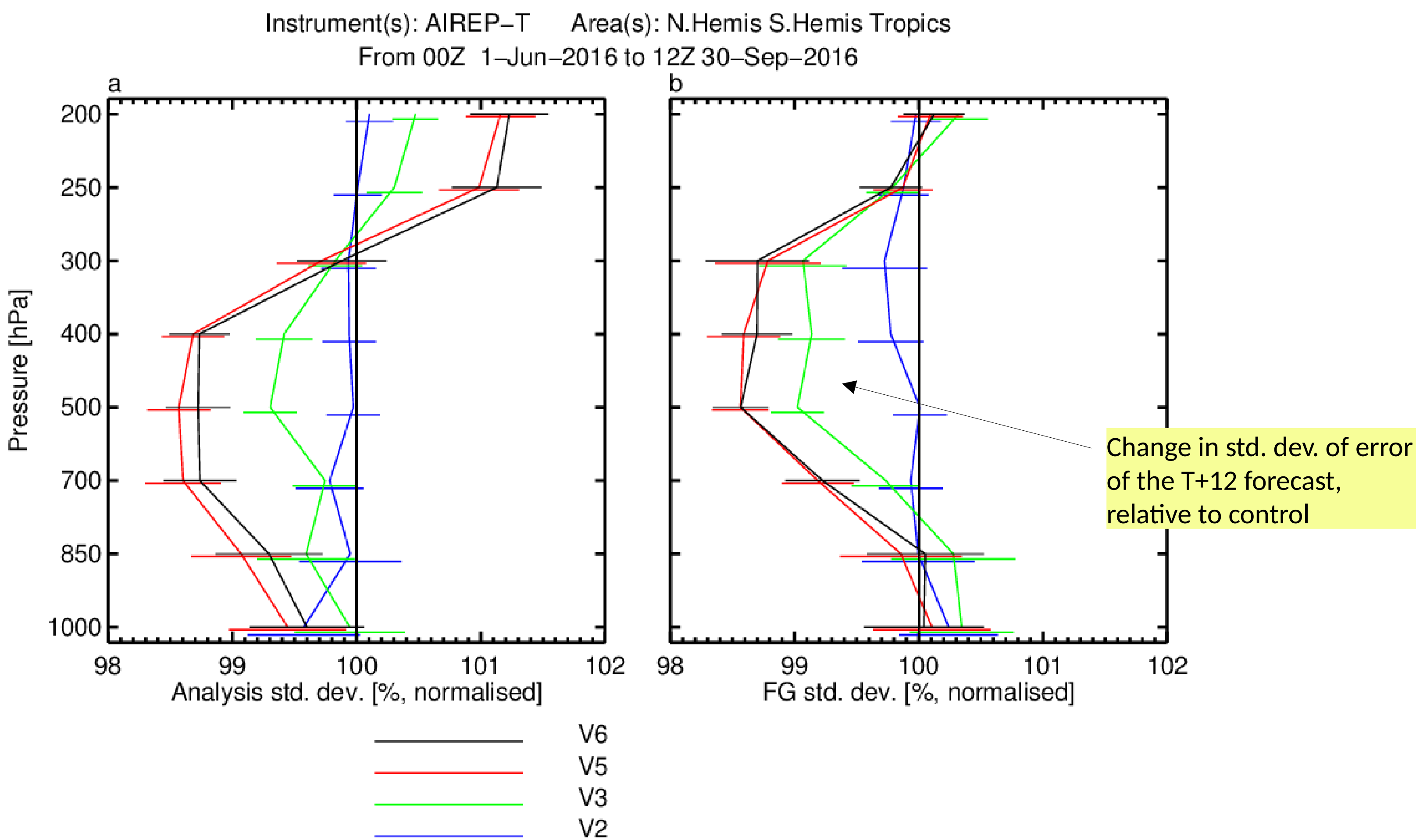
4. Are our scores meaningful? Changing the reference changes the results

Problem areas: Tropics, stratosphere, any short-range verification, any verification of humidity



Observational verification “obstats”

Example: verification against aircraft temperature measurements (AIREP)



Summary: four issues in operational R&D verification

1. Type I error due to multiple comparisons:
   - Try to determine how many independent tests *n* are being made (e.g. compute correlation between scores)
     - Paired differences in medium range dynamical tropospheric scores are all quite correlated
     - Paired differences are correlated at different forecast ranges
   - Once *n* is estimated, use a Šidák correction

2. Type I error due to time-correlated forecast error:
   - Chaos experiment used to validate an AR(2) model for correcting time-correlations
   - Note that at forecast day 10, this may not work: long-range time-correlations?

ECMWF

Summary: four issues in operational R&D verification

3. Type II error because typical experiments test only small changes in forecast error:
   - 300-400 forecasts are now a minimum requirement for research experiments at ECMWF

4. Are the forecast scores meaningful?
   - Own-analysis scores are accurate in the medium and long-range, for midlatitude dynamical scores
   - In other areas (e.g. tropics, stratosphere, early forecast range) these scores are often measuring something very different from forecast skill
     - Also check observational-based verification

For more detail on issues 1-3 see Geer (2016,Tellus) “Significance of changes in forecast scores”