“Twin-analysis” verification: a new verification approach that alleviates pitfalls of "own-analysis" verification when applied to short-range forecasts

Daisuke Hotta[1], Takashi Kadowaki[2], Hitoshi Yonehara[2], Toshiyuki Ishibashi[1]

[1] MRJ/JMA, [2] NPD/JMA

Standard forecast verification practice at operational NWP centers

Verif. against obs.

Pros

- Forecast errors and observations errors can be reasonably assumed independent

Cons

- Limited/sparse spatial coverage
- Intricate data handling

Verif. against "own-analysis"

Pros

- Uniform spatial coverage
- Ease of data handling

Cons

- Forecast errors and analysis errors (with respect to the (unknown) truth) can be positively correlated

→ Can result in overly optimistic scores

Issues with “own-analysis” verification:

- Positive correlation between forecast and analysis errors often makes interpretation difficult (counter-intuitive or even misleading).
- Algebraic explanation

$$\mathrm{RMSE}_{\text{vs-anl}}{}^2 = \mathbb{E}[(f\text{-}a)^2] = \mathbb{E}[(f\text{-}t)^2] + \mathbb{E}[(a\text{-}t)^2] - 2\mathrm{Cov}(f\text{-}t, a\text{-}t)$$

$$= \mathrm{RMSE}_{\text{true}}{}^2 + (\text{Anl RMSE})^2 - 2 * (\text{Error corr}) * (\text{Fcst RMSE}) * (\text{Anl RMSE})$$

where f: forecast, a: analysis, t: truth, $\mathbb{E}$: expectation over many cases

- Implication:
  - RMSE scores can be lowered if error correlation increases
  - even when true fcst error is unchanged (or even degraded).

Issues with “own-analysis” verification: Examples

- Feeding new observations to data-sparse regions induces apparent “forecast degradation” despite improvement in O-B fits (e.g., Bouttier and Kelly, 2001).
- Re-using information from the first guess (such as in retrieval assimilation) can apparently “improve” scores (which is overly optimistic) (e.g., Geer et al. 2010 Part II).
  - Extreme example: Forecast-forecast cycle (i.e., assimilating no observations at all) gives perfect score (i.e., RMSE=0)
- → Extra-caution is necessary when interpreting “own-analysis” verification, particularly for short-range forecasts.

Sources of positive correlation between forecast and analysis errors

- (1) forecast and analysis sharing the same "ancestry"
  - The impact stronger for shorter lead times
  - stronger also when the observational information is less incorporated in the analysis, e.g.,
    - when observation error variance (**R**) is large
    - or when fewer observations are assimilated
- (2) forecast and analysis sharing the same bias
  - due to the use of the same forecast model

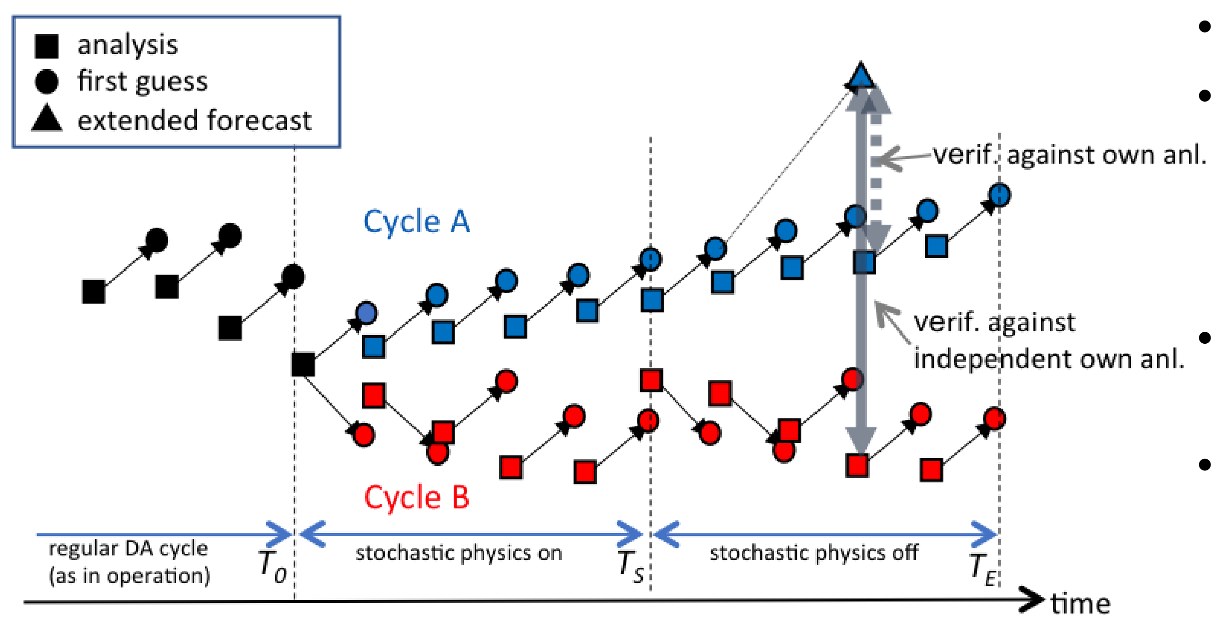
- The bias issue (2) is very difficult to tackle.
- In this study we focus on (1) and try to isolate the random component of the correlation term $-2\mathrm{Cov}(f-t,a-t)$

Proposal for a new verification method: "Twin-analysis verification"

- $\mathrm{RMSE}_{\text{vs-anl}}{}^2 = \mathbb{E}[(f\text{-}a)^2] = \mathbb{E}[(f\text{-}t)^2] + \mathbb{E}[(a\text{-}t)^2] - 2\mathrm{Cov}(f\text{-}t, a\text{-}t)$
- We wish to isolate the contamination from the term $- 2\mathrm{Cov}(f\text{-}t, a\text{-}t)$
- How? → Verify against an independent realization *a′* of analysis that follow the same probability distribution as that of the own analysis *a*
- How to generate the independent analysis *a′* ?
- → Employ "twin" cycle (Inspired by the approach of Kotsuki et al. (2019) for ensemble FSOI)
    - Use the same assimilation system assimilating the same set of observation
    - But initialize the cycle at a sufficiently earlier time from an independent first guess
    - which is generated by switching on stochastic physics

Graphically
explained i
the next
slide

Experimental set-up



spin-up period:

- The cycle stochastically splits into two.
- Both cycles run with stochastic physics turned on.

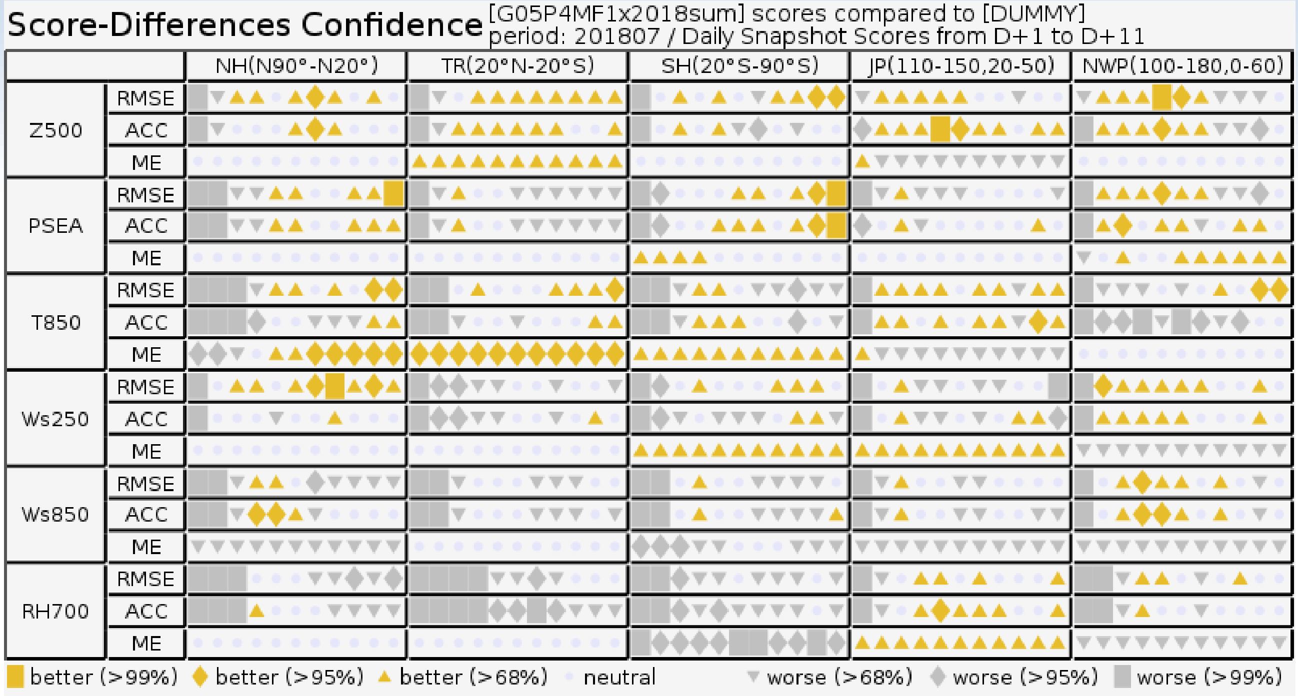
verification period:

- Both cycles now run with stochastic physics turned off.
- Extended forecasts from cycle A are verified against either cycle-A analyses or cycle-B analyses.

- Using the operational 4DVar,
- Initialize Cycles A and B from two independent analyses that can be considered drawn from the same distribution
- using the same model and observations
- so that their bias tendency should be equivalent.

- Compare the scores of
  - Cycle A fcst verified against Cycle A analysis (CNTL), and
  - Cycle A fcst verified against Cycle B analysis (TEST)
- **Discrepancy between TEST and CNTL is an indication of contamination from the correlation term** **$-2\mathrm{Cov}(f\text{-}t, a\text{-}t)$**

Results: Score differences and their statistical significance

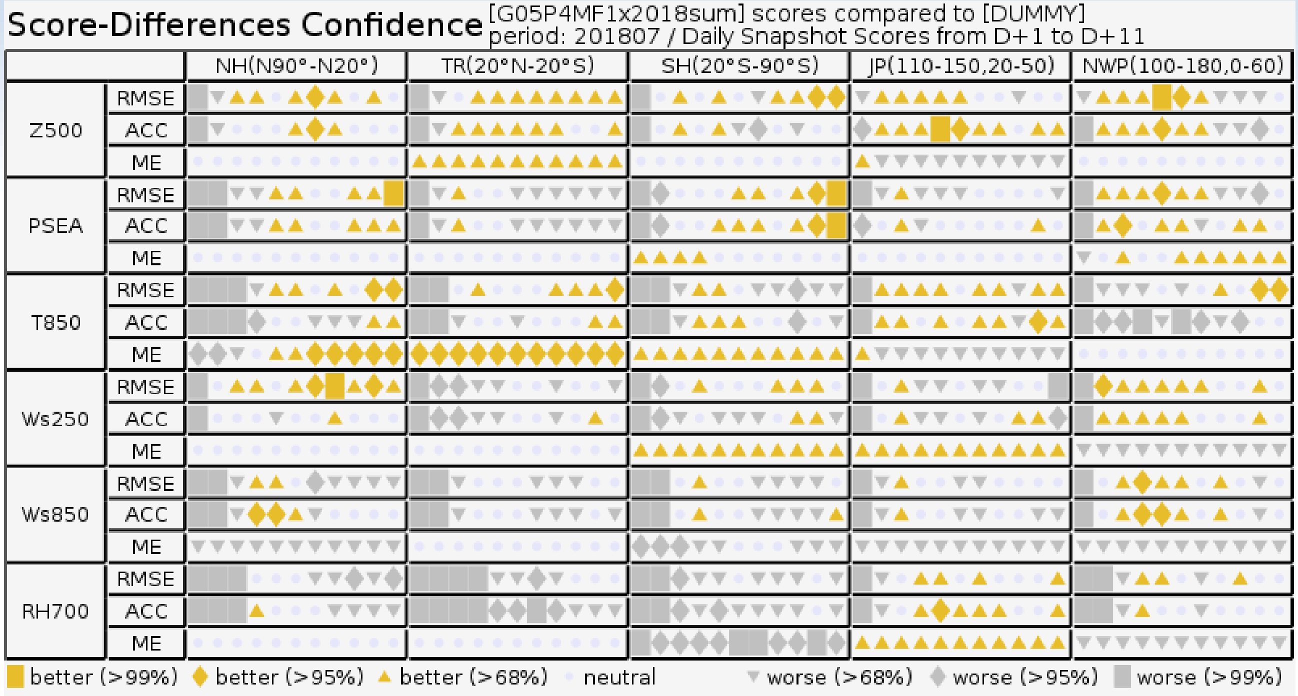


Comparison of "own-analysis" (CNTL) and "twin-analysis" (TEST) verification scores computed for the same forecast.

Note that any difference in the scores are just "artefacts" that arise from difference in verification methodologies

- For any elements and any areas, both RMSE and ACC scores exhibit statistically significant "degradations" for short lead times (up to ~ 2 days)
  - which highlights the over-optimism of "own-analysis" verification
- RMSE and ACC scores are quite consistent
- The "longevity" of score differences varies depending on the verified elements and regions
  - Z500 and Ws250 (wind speed at 250hPa): up to only ~ 1 day
  - T850 and RH700: persists up to ~3 days and beyond

Results: Score differences and their statistical significance

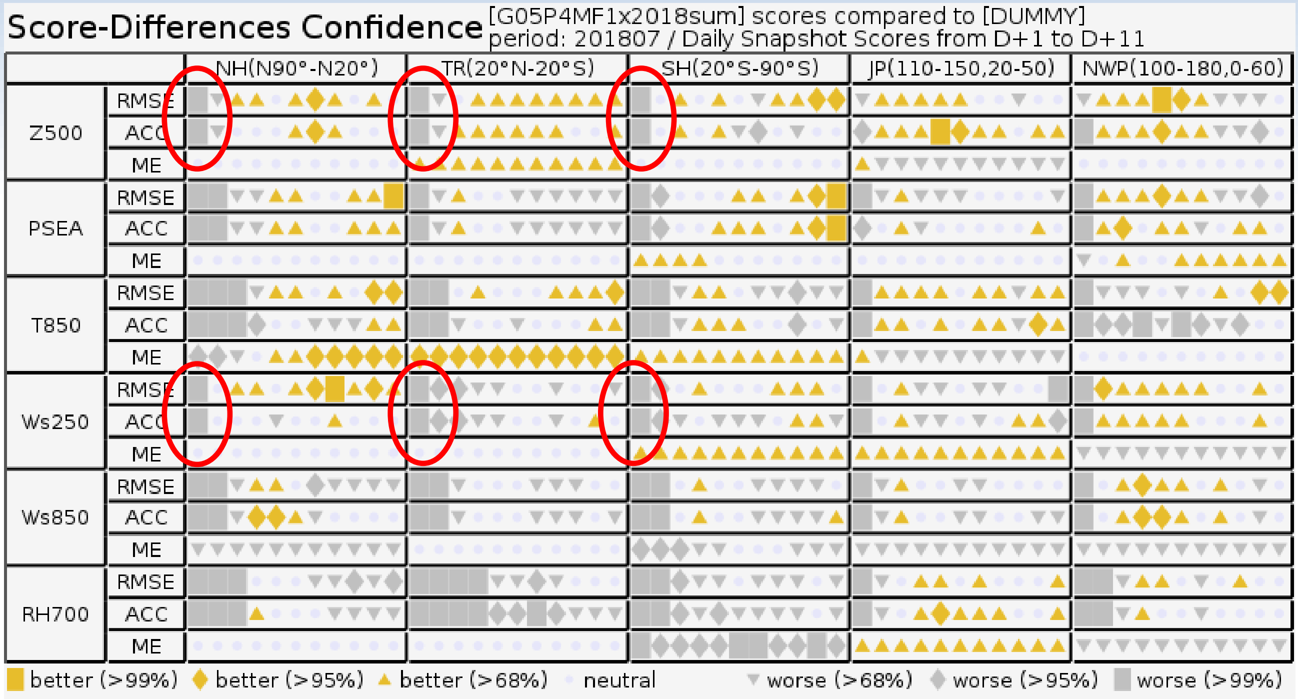


Comparison of "own-analysis" (CNTL) and "twin-analysis" (TEST) verification scores computed for the same forecast.

Note that any difference in the scores are just "artefacts" that arise from difference in verification methodologies

- For any elements and any areas, both RMSE and ACC scores exhibit statistically significant "degradations" for short lead times (up to ~ 2 days)
  - which highlights the over-optimism of "own-analysis" verification
- RMSE and ACC scores are quite consistent
- The "longevity" of score differences varies depending on the verified elements and regions
  - Z500 and Ws250 (wind speed at 250hPa): up to only ~ 1 day
  - T850 and RH700: persists up to ~3 days and beyond

Results: Score differences and their statistical significance

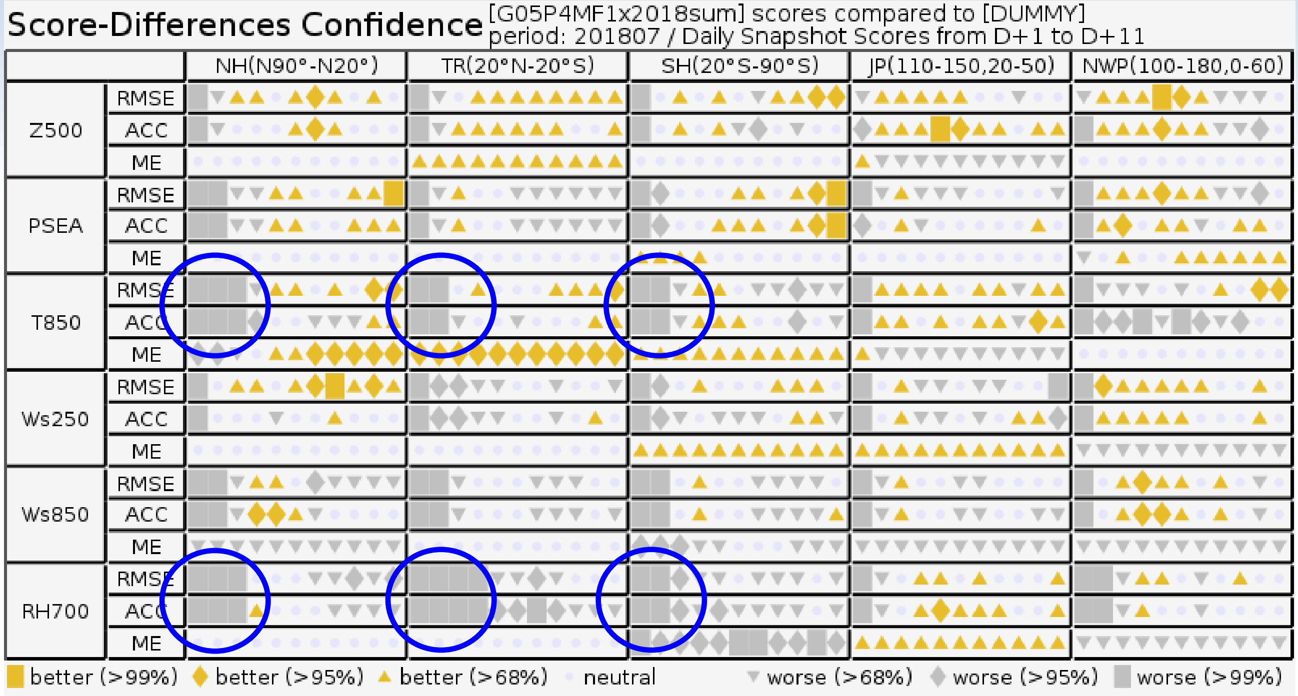


Comparison of “own-analysis” (CNTL) and “twin-analysis” (TEST) verification scores computed for the same forecast.

Note that any difference in the scores are just “artefacts” that arise from difference in verification methodologies

- For any elements and any areas, both RMSE and ACC scores exhibit statistically significant “degradations” for short lead times (up to ~ 2 days)
  - which highlights the over-optimism of “own-analysis” verification
- RMSE and ACC scores are quite consistent
- The ”longevity” of score differences varies depending on the verified elements and regions
  - Z500 and Ws250 (wind speed at 250hPa): up to only ~ 1 day
  - T850 and RH700: persists up to ~ 3 days and beyond

Results: Score differences and their statistical significance

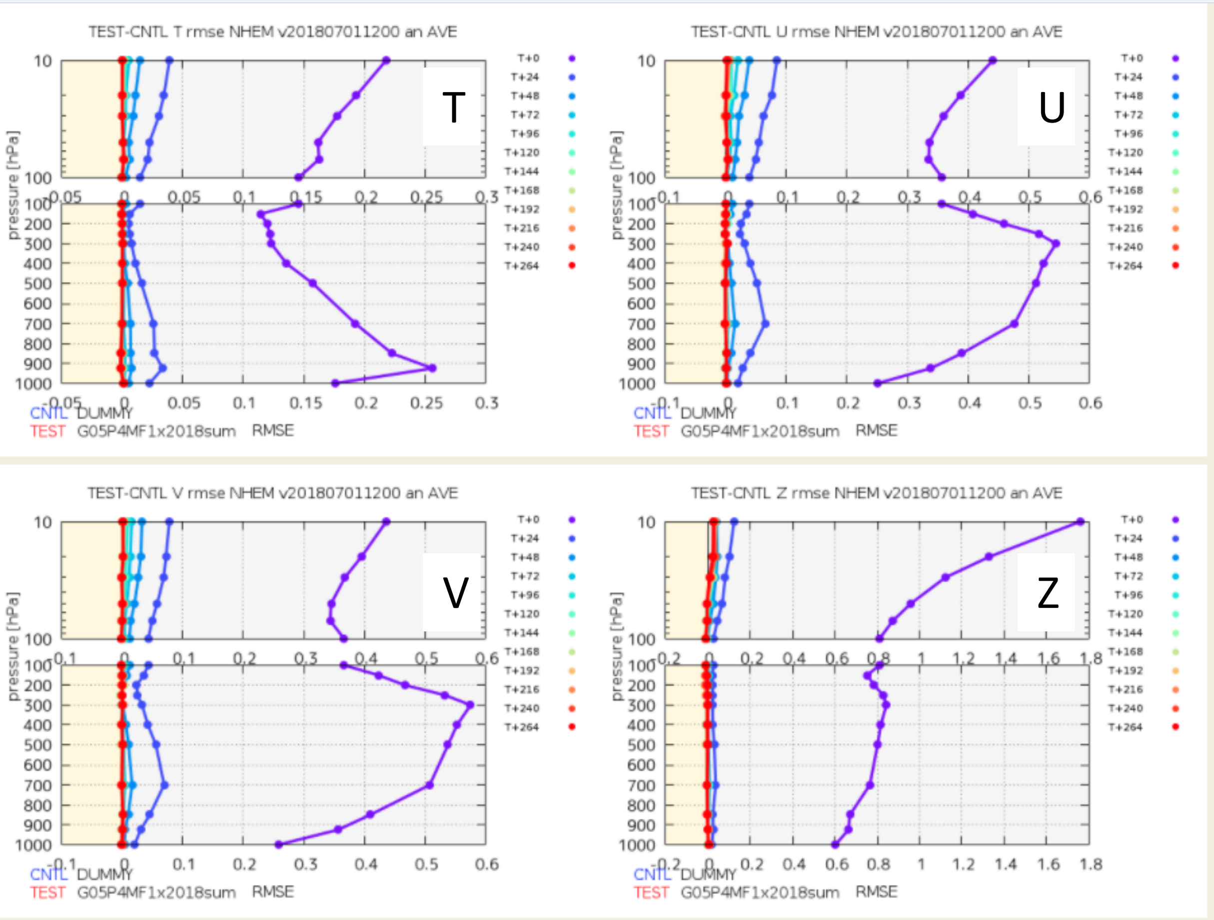


Comparison of "own-analysis" (CNTL) and "twin-analysis" (TEST) verification scores computed for the same forecast.

Note that any difference in the scores are just "artefacts" that arise from difference in verification methodologies

- For any elements and any areas, both RMSE and ACC scores exhibit statistically significant "degradations" for short lead times (up to ~ 2 days)
  - which highlights the over-optimism of "own-analysis" verification
- RMSE and ACC scores are quite consistent
- The "longevity" of score differences varies depending on the verified elements and regions
  - Z500 and Ws250 (wind speed at 250hPa): up to only ~ 1 day
  - T850 and RH700: persists up to ~ 3 days and beyond

Results: Vertical profiles of RMSE score differences NH extra-tropics (similar in SH extra-tropics)



Score differences at T+0 is the RMS diff. between the “twin” analyses

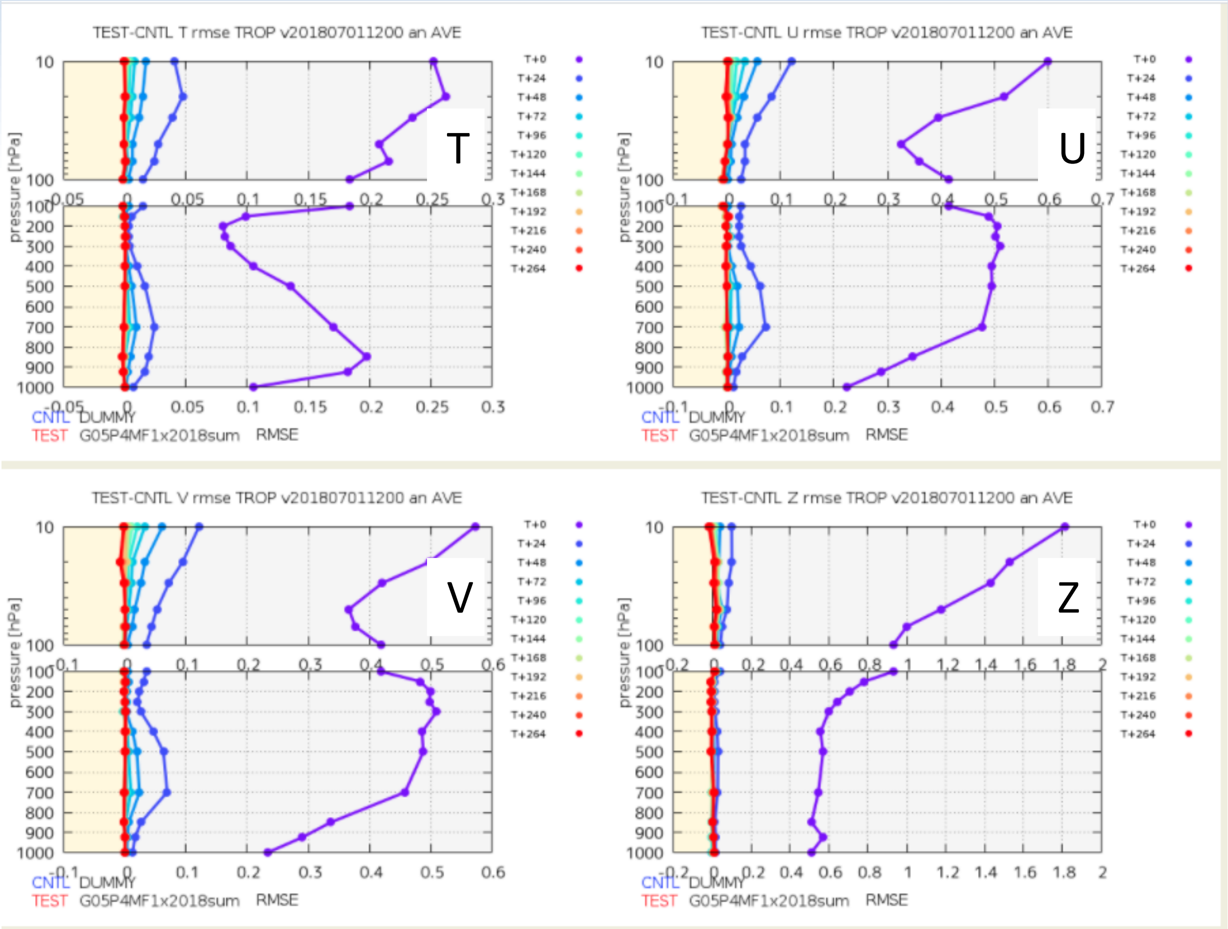
→ Can be interpreted as an indication of to what extent observations can constrain the analysis uncertainty

Large discrepancies found in

- Temperature at lower troposphere and upper stratosphere
- Winds at mid-to-high troposphere and upper stratosphere
- Height field at upper stratosphere

Coincides with regions where obs. are scarce

Results: Vertical profiles of RMSE score differences Tropics



Similar to the NH and SH extra-tropics, but the differences persist to longer lead times in the upper stratosphere (again data-sparse region)

Summary

- The “own-analysis” verification scores can be unreliably optimistic at short rages
  - due to the error correlation between forecast and analysis
- “Twin-analysis” verification is proposed and conducted to quantify to what extent “own-analysis” scores are contaminated by the error correlation.
- Results suggest that:
  - Spurious optimism persists at least 1 day
  - can persist up to 3+ days for some elements and regions
- The spurious effect (= uncertainty of “own-analysis” scores) persists longer for relatively unobserved regions and elements

Implications

- The difference between “twin-” and “own-” analysis scores can be interpreted as the uncertainty of “own-analysis” scores

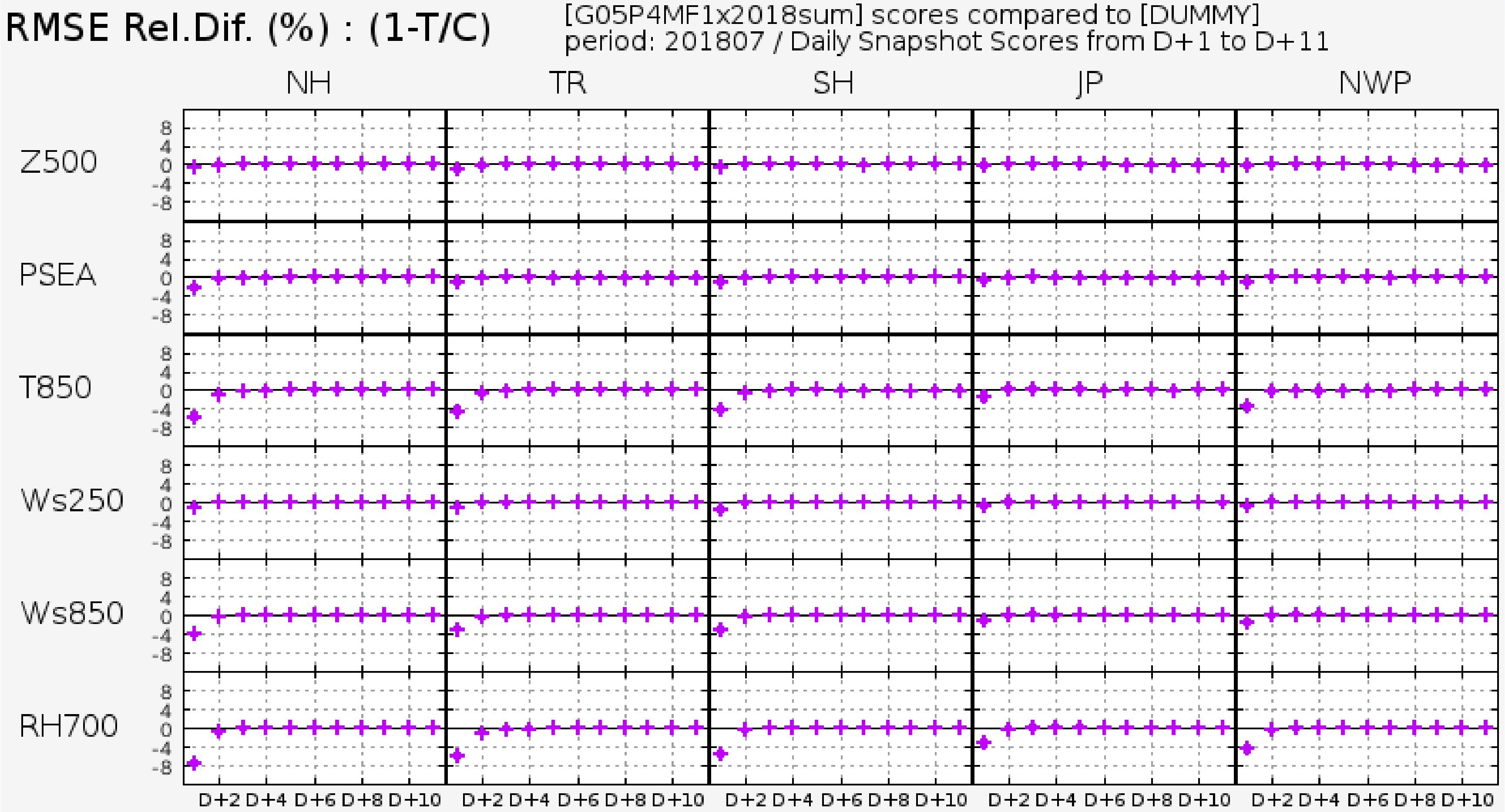
  → perhaps can be used to estimate the reliability of the scores (like a confidence interval)
- From our experiments, the difference between the scores was quite large
  - for Z500 T+24 score, the difference was comparable to using or not using an AMSU-A instrument
- Practical recommendation (maybe controversial):
  - Ignore degradations in short-range own-analysis scores (up to ~ 1day)

References

- Bouttier, F and G. Kelly (2001) Observing-system experiments in the ECMWF 4D-Var data assimilation system. *Q. J. R. Meteorol. Soc.,* **127,** 1469–1488.
- Geer, A.J., P. Bauer and P. Lopez (2010) Direct 4D-Var assimilation of all-sky radiances. Part II: Assessment. *Q. J. R. Meteorol. Soc*., **136**, 1886–1905.
- Kotsuki, S., K. Kurosawa and T. Miyoshi (2019) On the properties of ensemble forecast sensitivity to observations. . *Q. J. R. Meteorol. Soc*., **145**, 1897–1914. doi: 10.1002/qj.3534

BACKUP

RMSE score normalized difference



Anomaly correlation score normalized difference