Response to Simon Zwieback's interactive comment on "Multi-scale analysis of bias correction of soil moisture"

We thank our colleague Simon Zwieback for his interest and comments on our manuscript. We take this opportunity to consider his invaluable comments (copied here and identified by [SZ]) and improve the clarity of this work. In particular we extended our discussions to reflect upon the insights provided by our colleague. Our responses are identified by [A] while extracts of the specific changes made in the revised manuscript are shown in quotations and in blue.

[SZ] The authors present a framework within which soil moisture time series (as derived from e.g. models or remote sensing instruments) can be analysed and compared at different temporal scales. Such data commonly exhibit complex scale-dependent behaviour: a fact to which only cursory attention is usually paid when soil moisture products are assessed or compared. The manuscript is thus certainly relevant for HESS - and the hydrological community at large. I also find it well written and generally carefully argued, but I would like to mention a few points that the authors might want to consider:
[A] Indeed our work attempts to address this gap by describing a more systematic approach (wavelet-based multi-resolution analysis in the temporal domain) to analyse the time scale-dependent variability between different data sets such as of soil moisture variable. It is also prospective to consider multi-resolution analysis in the spatial domain, but at this stage, multiple independently-derived high-resolution soil moisture data sets (e.g., from Sentinel satellites) are yet to be available for comparisons.

[SZ] 1. Previous work
p 8998, 1-12: this is mostly based on hydrological principles, previous empirical work (e.g. [1], [2], [3]) not being mentioned
[A] We agree that an alternative approach undertaken by Loew and Schlenz (2011) [1], Su et al. (2014) [2] and Zwieback et al. (2013) [3] is to consider (moving) windowed statistics. In these approaches, the underlying assumptions are that the biases or errors have somewhat seasonal characteristics. We amended the manuscript to read,
"One possible remedy is to apply bias correction, either TC or statistical-moment matching, only to anomaly timeseries (Miralles et al., 2011; Liu et al., 2012; Su et al., 2014), but it remains unclear how these methods affect the signal and noise components in the corrected data. **Alternatively a moving window can be used to examine the time-varying statistics of timeseries (Loew and Schlenz , 2011; Zwieback et al., 2013; Su et al., 2014).**"

[SZ] 2. Interpolation and interpretation of the results
p 9000, 1-2: how sensitive are the results to the choice of interpolation algorithm? I would expect it to be particularly relevant at fine temporal scales, but this is not included in the analysis of Section 4 (e.g. lines 14-15 on p 9004).
[A] This is a legitimate concern, especially for AMSR-E that had a revisit time of 1-2 day and limited sensor swath. The different influence of different interpolation algorithms will be most apparent over extended gaps. We show below the relative frequency of gaps of different lengths (1/2-day, ≤1-day, ≤2-day). Over 95% of the gaps in AMSR-E data at most regions of Australia have lengths of 1-day or less (b). By contrast, most of the gaps in the *in situ* data are considerably much longer but infrequent and the interpolated values were not included in statistical analyses.

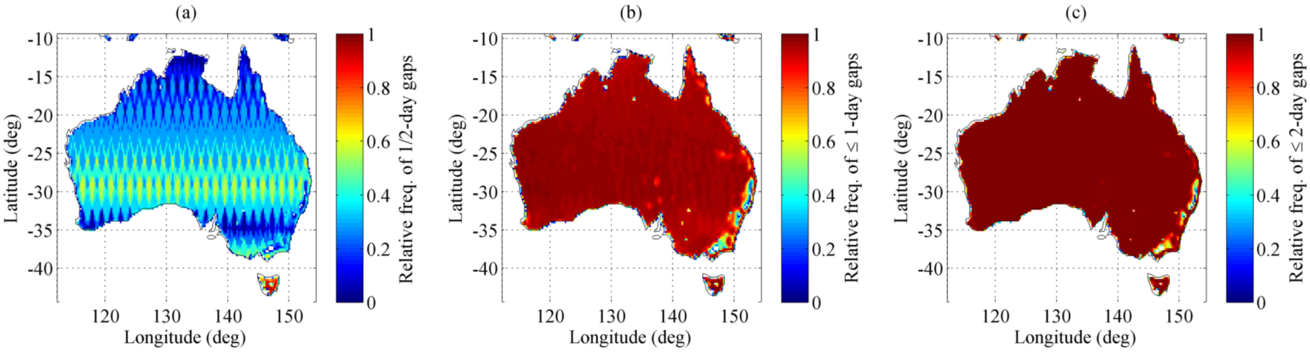

Figure 1: Analysis of the length of gaps in the AMSR-E data over Australia.

Hence we focus on the gap-filling of the AMSR-E data. Below we compare four interpolation algorithms, namely discrete cosine transform (DCT) algorithm reported in Wang et al. (2012) [cited in

the paper], nearest-neighbour, linear interpolation, and piecewise spline, and they were applied to gaps of length ≤5 days. Note that DCT and the 5-day threshold were adopted in our study. We observed that DCT interpolated data show greatest similarity with linear interpolation, largely due to the short lengths of the gaps and the frequent occurrence of gaps. The nearest-neighbour interpolation is expected to introduce more errors to the data, while cubic spline interpolation algorithm is observed to produce spurious peaks. While we expect our results are sensitive to the choice of the interpolation methods, we argue that DCT and linear interpolation are better methods to use for AMSR-E data.

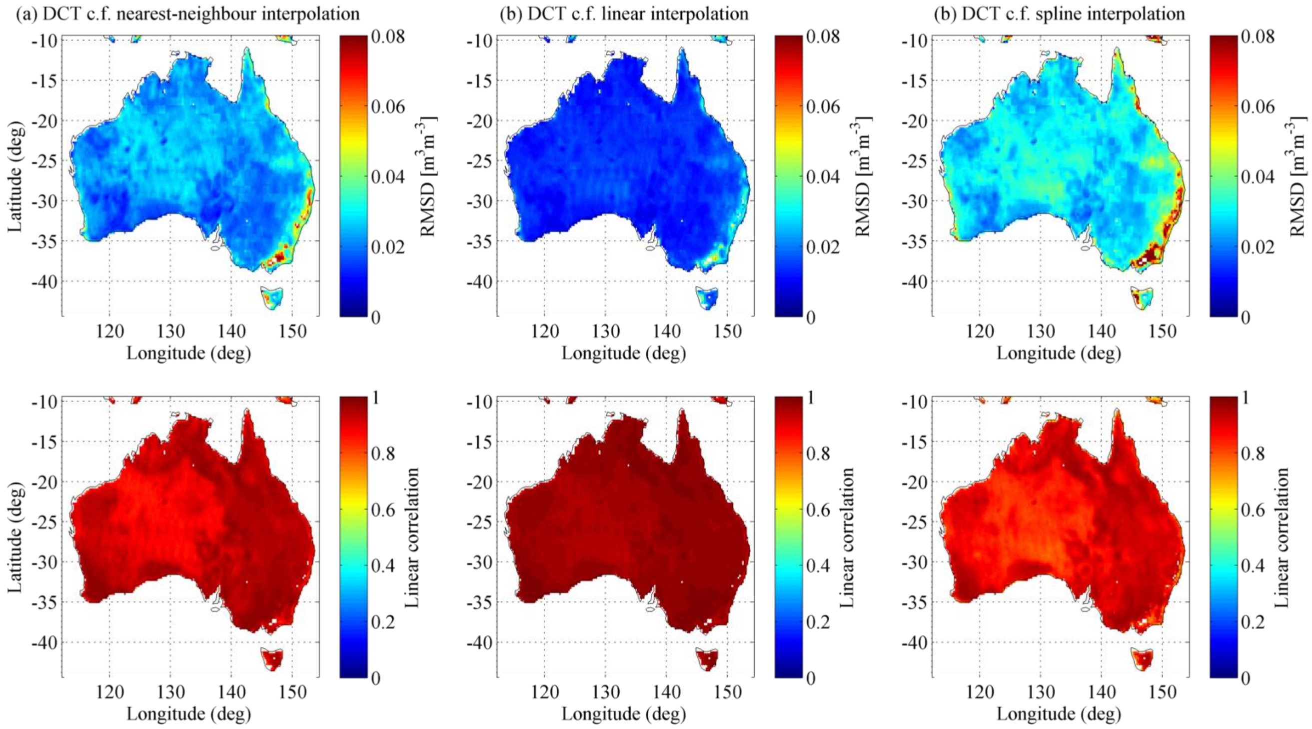

Figure 2: Comparisons of discrete cosine transform (DCT) based interpolation method (Wang et al., 2012) and traditional methods, namely nearest-neighbour, linear and spline interpolations. The differences between DCT and the traditional methods are quantified using root-mean-square difference (RMSD) and Pearson's linear correlation.

To make note of this consideration, we amended the manuscript with the following text,

"For use in wavelet analysis (Sect. 4), a one-dimensional (1-D in time) interpolation algorithm (Garcia, 2010) based on discrete cosine transform (DCT) (Wang et al., 2012) was applied to infill gaps of lengths ≤ 5 days in AMSR-E. **Other interpolation methods were trialled; e.g., linear interpolated AMSR-E shows great similarities to the DCT interpolated data while cubic spline interpolation leads to spurious peaks."**

[SZ] More generally, the whole discussion seems to be based on a model that can represent discrepancies between two soil moisture products by noise and multiplicative biases, which has not been introduced at that point. I think that the section, and similarly Sec. 5, would be improved by clarifying this aspect, as well as by considering different descriptions of the discrepancies, as the assumption of temporal stationarity at any scale seems to be not easily tenable (e.g. apparent presence of secular trends).

[A] Indeed in this section we adopt the viewpoint that the correlations between different data are diminished by the presence of noise, while differences in spread (i.e., Δstd) are influenced by noise as well as multiplicative bias. Of course, presences of extraneous signals and nonlinearity will also influence the observed correlations and Δstd. Our adopted model is therefore a simplification, and there is a need to also highlight its limitation in the paper. We added the following text to Section 4 to reflect these.

"… we recall that weak R indicate presence of noise and/or presence of nonlinear correlation between any pairs of the data, while differences in standard deviation (Δstd) can also indicate presence of noise, extraneous signal and/or multiplicative bias. Typically one invokes a linearity assumption and assumes an affine relation between the signal components of the different data and an additive noise model (more later in Section 5), so that these differences between the data are attributed to an overall additive bias $E(X) – E(Y)$, multiplicative biases, and noise. While we adopt this simplistic viewpoint here, its limitations to properly account for variable lateral and vertical measurement supports should be noted. For instance, short-time scale SM dynamics shows increasingly attenuated in amplitude but also delayed in time in deeper soil columns (e.g., Steelman et al., 2012). Additionally SM is physically bounded between field capacity and residual content and these thresholds can vary with soil texture, location and depths. These effects can give rise to temporal autocorrelation in errors and undermine the linearity assumption between coincident measures. Finally, the non-stationary characteristic of noise in satellite SM (Loew and Schlenz , 2011; Zwieback et al., 2013; Su et al., 2014a) due to e.g., seasonal dynamical land surface characteristics such as soil moisture (Su et al., 2014b), is not treated here."

These trends, as well as more general additive biasses such as seasonal variations, could also furnish a parsimonious description for the discrepancies between products, e.g in Fig. 8a) in [4]; so would autocorrelated noise, the two being quite closely related [3]. They might not be easily incorporated into the framework, but by virtue of this, the analysis of such cases could aid future interpretation of data within this framework: how would, for instance, a seasonal additive bias be represented if such data were analysed with this model? These issues are only briefly touched upon in the conclusions.

[A] We agree with our colleague. The perception of discrepancies between any two time series can vary, depending on the time-scale of the analysis. For instance, a short time window can be used to observe temporally varying additive bias. Using a long time window, such additive bias may manifest as multiplicative bias, e.g. differences in the amplitudes of the seasonal signal in the two data. This equivocal definition of additive or multiplicative bias is inherently a scale issue. The strength of wavelet analysis is decomposing a timeseries of no mean (additive) bias into multiple (with different frequency) timeseries with no mean additive bias, but only with multiplicative bias. At a given scale j, because the detail timeseries p_j does not contain variations of time scale $>j$, the weak-sense stationarity conditions for TC analysis with long timeseries can be better satisfied.

Further, often subjective adoption of different signal and noise models may lead one to interpret the multiplicative bias as autocorrelated noise; e.g., a coincident signal model c.f. a non-coincident signal model, and the presence of extraneous signal unique to one data.

In sum, the chosen time-scale of analysis and chosen signal/noise model therefore influence our interpretation of the discrepancy between data. In our work (by using the MRA model in Eqs. 4-5 and near-decade long time-scale analysis), we assume that time-varying additive bias manifests as an overall multiplicative bias, and we also assumed coincident signal model, orthogonal error, cross-correlated error, and absence of extraneous signal. The latter three assumptions are often used in triple collocation analysis of soil moisture. Of course, these assumptions are yet rigorously tested, until more recently by Yilmaz and Crow (2014). These viewpoints are now added to the manuscript in Section 5.

"…However in this work, we consider only variability across j and assume stationarity at each scale. Pearson's linear correlation R and variance analyses (see Appendix A) are performed on the Kyeamba's INS, AMS and MER SM (as p in Eq. 2) detail (p_j) and approximation ($p^{(a)}_j$) timeseries in Fig. 3. **The strength of MRA is that since the detail timeseries p_j at a given scale j does not contain variations at time scales $> j$, the weak-sense stationarity conditions can be better met.**

…

The interpretation of the discrepancies between X and Y can vary depending on the time period of the data and the analysis, and the adopted signal/noise model. By using entire 9-year record of INS, AMS and MER data in MRA, the MS model does not observe time-varying additive bias (e.g., from using the moving-window approach (Su et al., 2014a)) and autocorrelated errors (from using lagged covariance (Zwieback et al., 2013)). **Rather, MRA and the MS model enable a description of the systematic differences to be wholly based in terms of multiplicative biases at individual time scales, and the random differences in terms of additive noise**. Specifically, this contrasts with the short time-window approach (≤ 32d), where multiplicative bias existing at coarse scales (e.g., $p^{(a)}_6$) will manifest as both time-varying additive and multiplicative biases.

…

The standard assumptions of orthogonal and mutually uncorrelated errors are used, so that the covariance $\mathrm{cov}(f_j, \varepsilon_{p,j}) = 0$, $\mathrm{cov}(f', \varepsilon'_p) = 0$, $\mathrm{cov}\,\varepsilon_{p,j}, \varepsilon_{q,j}) = 0$, $\mathrm{cov}(\varepsilon_{p,j}, \varepsilon'_q) = 0$ and $\mathrm{cov}(\varepsilon'_p, \varepsilon'_q) = 0$ for $p \neq q$, $p,q \in \{X,Y\}$."

[SZ] 3. Definition of model and relevant quantities

Sec. 5: which quantities are random and which are deterministic? If the time series are assumed to be realizations of stochastic processes (what kind of expectations are understood by the operator E?), which properties are attributed to these stochastic processes, esp. with regards to the wavelet representations, cf. [5] but also Appendix A, where they seem to be treated as deterministic. Are E(p) and E(f) time-variant?

[A] From a measurement and sensor point of view, f is a deterministic signal such as soil moisture, but p is stochastic due to the random nature of measurement noise from radiometric inaccuracy or background contamination, etc. Note that serially uncorrelated noise (as assumed in our model) will be represented by serially uncorrelated coefficients in wavelet domain. Hence from data inter-comparison

viewpoint then, *f* and its associated wavelet coefficients are interpreted as deterministic. By contrast, *p* and its associated wavelet coefficients are interpreted as probabilistic.

This should however be distinguished from a physical viewpoint: *f* is a single physical realisation of the stochastic process (soil moisture is driven by stochastic forcing from rainfall, plant absorption, solar radiance/land surface temperature fluctuations). From the MRA, *f* contains high to low frequency components, and it is our viewpoint that all the components of *f* are stochastic. As we only have a single realization of the process, the statistical properties of the process can only be inferred from the statistics of *p* and *f*.

[SZ] 4. Error structure

p 9010, 8-20: you present the modification of the error-structure by scale-dependent bias correction as an unwelcome side effect. I do not think this is necessarily the case: it depends on which representation/transformation of the time series one is primarily interested in. As the careful analysis of diverse patterns of soil moisture time series is a great asset of this manuscript, I would welcome a slightly more detailed discussion.

[A] We agree with our colleague. Our focus was the representation of the timeseries and the error on the whole after reconstruction. If the focus was one of the detail timeseries, one may not worry about the amplification of the error as the associated signal-to-noise ratio remains unchanged after linear rescaling. Furthermore, in response to the question whether the modification of the error structure is desirable, it depends on the specific use of the bias-corrected data. We revised the text to highlight our desirable outcome of the bias correction:

"On the other hand, the A/S-based and MS methods can modify the original error profiles in the data across the scales, by amplifying (or suppressing) errors in individual components (either Y_j , Y_S, or Y_A) with less-than (greater-than) unity pre-correction α's. **This may be considered undesirable for an objective to produce more physically representative data with a simple error structure on the whole.** Therefore arguably, none of these methods is entirely satisfactory, in manners of not removing the multiplicative biases completely and/or changing error characteristics. **From this viewpoint,** the task of bias correction is seen as inseparable from that of noise reduction when considering MS (or A/S) bias correction, unless certain components in MRA were explicitly ignored.

…

The last example presents an impetus to consider noise removal prior to bias correction **and produce a simpler error structure in the bias corrected data $Y*$.**"

Please also refer to our response to the comment about the chosen optimality criterion below.

5. Minor points

[SZ] p 9001: please clarify the meaning of j, j0, and J: N = 2j , but then it seems to be 2J

[A] This is a typographical mistake. It should read "$N=2^J$".

[SZ] p 9002: is the (evenly sampled) time *t* dimensionless or not? The temporal location of ϕ_{jk} is stated as $k*2^j$, which is dimensionless.

[A] For clarity, we include a term Δt to represent the sampling interval of the timeseries and rewrote the text as follows,

"The 1-D orthogonal discrete wavelet transform (DWT) enables MRA of a timeseries $p(t)$ of dyadic length $N=2^J$ **and a regular sampling interval Δt** by providing

…

with **scale of variability $2^j\Delta t$ and temporal location $k2^j\Delta t$.** The weighting or wavelet coefficients,…"

[SZ] p 9002, 23: the significance being based on what test and significance level?

[A] The analysis aims to illustrate that the trends in the three data show differences, in particular in terms of their gradients. We adopt the simplest method of fitting (using least-square) a linear trend line to the coarsest approximation time series, and statistical testing was not conducted to test for the significance of the trend. To clarify, we revised the text as follows.

"…Fitting a trend line to their coarsest scale approximation series suggests that the trends (magneta lines) in the three data show different gradients, with the trend in INS showing the smallest positive gradient. The differences in dynamic ranges of their detail and approximation timeseries, together with their mismatch in shape and trend, are indicative of multiplicative biases and noise. **…"**

[SZ] p 9005, 23: that is rather consistency (and it is a limit in probability)

[A] Yes, to clarify this, we revised the text slightly as follows,

“Within the operating assumptions of TC, TC estimates are unbiased **and consistent**; that is, the estimated $\hat{\alpha}_{Y,j} = \alpha_{Y,j}$ **as the asymptotic limit**.**”**

[SZ] p 9006, 14: is not the identity of the signal components (treated as a deterministic or random variable) the criterion by which optimality (or ideality) is defined?

p 9006, 14: different justifications for the estimation of α have been provided (consistent estimation of the slope between signal and measurement; matching of the magnitude of the signal component; orthogonality principle based on LMSE estimation, etc.). They depend on i) what one wants to actually estimate and ii) whether the signal component is treated as a random variable or a deterministic one. Which point of view is adopted in the manuscript?

[A] This is a very good point. The criterion for a matching Y to X depends on some choice of the optimality criterion. Here we define our optimality criterion based on matching the first two moments of the *signal* components in X and Y. As pointed out, this contrasts with matching the statistics of X and Y, and minimizing the differences between Y* and X. To make clear this viewpoint, we revised the manuscript as follows:

“Consider now the bias correction of *Y* to produce a corrected data *Y** that “matches” *X*. Different interpretations of a “match” and assumptions about signal and noise statistics lead to different bias correction schemes. **To describe matching, there are different choices of optimality criterion. First is based on matching the statistics of the signal-only component of *Y** to that of *X*. This approach requires consistent estimate of slope parameters α’s and the resultant statistics of *Y** and *X* may differ due to different noise statistics. Second is the based on the matching of the statistical moments between *Y**and *X* (e.g., VAR matching), although the statistics of their constitutive signal components may differ for the same reason. Third is based on the minimum-variance principle of minimizing the least-square difference between *Y**and *X* (i.e., the OLS estimation), but as already noted the estimator becomes inconsistent when there are measurement errors in *X* and *Y*.**

Here we define our optimality criterion based on the first criterion of matching the first two moments of the signal components in *X* and *Y* so that *Y*∗ is suitable for bias-free data assimilation. In particular, Yilmaz and Crow (2013) have shown that residual multiplicative biases due to sub- optimal bias correction scheme will cause filter innovations to contain residual signal and sub-optimal filter performance. Thus within the paradigm of the MS model, our goal of bias correction is to minimize the difference $|\alpha_{Y*,j} - 1|$ for $\alpha_{X,j} = 1$, so that the multiplicative bias terms in Eq. 6 are eliminated.”

[SZ] p9015, 5: what are physically meaningful results? There are many additional reasons why e.g. negative variances could be obtained, such as inadequate rescaling or cross-correlation.

[A] Indeed we require that the covariance are positive, while negative covariance can occur due to weak instrument (i.e., the signal components are too weak relative to the noise components), or cross correlation, or the inadequacy of the proposed affine signal and orthogonal error models. To make this point clear, we wrote the text as follows:

“When TC does not produced physically meaningful estimates **from negative covariance due to weak instruments and possible inadequacy of the considered signal and noise model**, the OLS estimator was used,

$$\hat{\alpha}^{OLS}_{Y,j} = \frac{\mathrm{cov}(X_j, Y_j)}{\mathrm{var}(X_j)} \quad \text{(B5)}$$

although its estimates are biased ($\hat{\alpha}^{OLS}_{Y,j} < \alpha_{Y,j}$) **for our purpose**, due to the extraneous contribution of noise variance in the denominator. **Similarly the VAR estimator can be used,** $\hat{\alpha}^{VAR}_{Y,j} = \frac{\mathrm{var}(Y_j)}{\mathrm{var}(X_j)}$**, but it is also biased.**”