Statistical modelling of collocation uncertainty in atmospheric thermodynamic profiles

Alessandro Fassò, Rosaria Ignaccolo, Fabio Madonna, and Belay B. Demoz

Answer to referee #2

We thank the reviewer for deepening the manuscript and rising a number of fundamental and stimulating questions. His/her comments have the potential to greatly improve the paper and we are trying to catch this opportunity.

As suggested by this referee, the main modification of our manuscript will be a new section before the methodological part of sections 2-4. As a motivating example, this section anticipates part of section 5, introduces the Beltsville case study and allows us to make the methodological part of sections 2-4 more easily readable and less cryptic.

Answers to numbered questions

1.1) It is not clear what the main objective of the paper is in the abstract and the introduction. Is it to determine the errors in a given atmospheric parameter profile?

In a sense the answer is yes, but much more at the same time. We are proposing a general methodology for understanding the uncertainty of those particular atmospheric profiles given by collocation error profiles. And an example for relative humidity is given.

1.2) …. Or is it to determine the co-location error between two atmospheric measurements?

Yes - as explained in the answer to the question 1.1.

1.3) … Or all of them at the same time?

In short, yes. Obviously, these three questions indicate to us that we need to improve on communicating the objective clearly.

To make this point clearer, we are going to improve the objective statements in the manuscript and in particular in the conclusions.

2) Please define adequately what are the definitions of the parameters being used in the paper. It is clear what an atmospheric observation is. But, what is an "environmental forcing factor"? The only field where I have heard this word is in climate science. Is it meant as other parameters which are measured at the same time as the main observed parameter under study? Or is it something else. Maybe a change in the name and a proper definition would help.

We are going to make this more clear in section 2 and 3. As requested, we will give a precise definition of environmental factors in general and we will extensively exemplify using radiosonde setup. The word “forcing” is meant to indicate a driving or controlling factor of the quantity in question, just as radiation difference is used as climate forcing. It is not uncommon to be used in statistics and set theory, way before it became fashionable in climate change discussions.

3) The mathematical notation in section 2 is very strange. I have never seen a function written as h->(s,t). Using a central dot instead of a variable also adds to the confusion.

We realize the paper straddles the area of statistical math and applied meteorological analysis and has resulted in notation inconsistency. We agree and we will modify the opening of section 2 and move to standard calculus notation for functions y(h), x(h) etc.

4.1) In the equation before Eq. 1, what is a "true smooth profile"?

The “true profile” is the atmospheric quantity (measurand) which is measured with its measurement uncertainty. In this manuscript the atmospheric quantity is assumed to be smooth, Hence the term “true smooth profile”.

4.2) Why does the true profile need to be smooth?

This point will be addressed in the revised manuscript. To begin with, note that the atmospheric quantities are assumed to be smooth but we do retain the information on errors. Our assumption, also allows for studying the variations that come from the degree of data smoothing.

Although the question of continuity is a big issue for physical quantities in general, this assumption is sensible at the space-time resolution we use for atmospheric variables. To see this, note that, in the literature on collocation, data are usually grouped and averaged on vertical bins of some 100-150m in depth, or more. This is the approach at the basis of formula S(h)^2, reported just before Eq. (8). In doing this a very low band smoothing is superimposed to data.

On the contrary, using our new approach, the vertical resolution of the data, which is usually in the range of 5-10m, is not substantially degraded in computing collocation errors thanks to the functional data approach. In fact we are able to compute the collocation mismatch of equation (9), namely μ(h)-μ°(h), in high resolution.

Note that, from the point of view of general functional analysis, it is not necessary to assume smoothness as is implied by the use of penalized splines as we do. Actually this approach can be extended to wavelets basis functions instead of splines in order to cover multiscale phenomena or other basis functions. But this is not the scope of the manuscript.In this manuscript, we suppose smooth atmospheric quantities. In fact we use penalized splines which try to balance fitting and smoothing by giving a penalty for large second derivatives.
Our approach is self-critical in the sense the measurement error is assessed in the uncertainty budget. This gives also an indication on how acceptable is the smoothing used.

5) In Eq. 1, Why is the true smooth profile a linear function of the forcing factors?

Generally speaking HRM is not a linear model for two reasons. First the last summand ω is a stochastic component, second β(h) is indeed a function, so the linear relationship depends on height. In other words it is locally linear plus an heteroskedastic error whose size and behaviour is quite important.

6) In Eq. 2. Why is the variance a linear function of the forcing factors?

Some of the observations to point 5) hold also here. This is a model for the irreducible uncertainty.
Note that both this assumptions are statistically assessed for the data at hand.

7) In Eq. 5, a variance is defined as U, which is confusing, because before the parameter was called u and the variance was denoted by sigma.

Agree. Notation will be uniformed.

8) Eq. 6.1. The variance is decomposed in three terms. Could you give a physical justification why this is so, besides citing a reference to the appendix? Could you explain more clearly what the different terms are? For example, is drift uncertainty the same as or related to co-location error? What is the environmental error? Does atmospheric turbulence, to cite one example, have anything to do with these errors?

Yes, turbulence in general is related to heteroskedasticity. This question is very interesting and could be the topic for an entire manuscript per se and would defining for which atmospheric variables this interpretation holds.
The title of this manuscript is "Statistical modelling of …" and Eq. 6 is a mathematical consequence of Eq.s (1) and (2).
The term U_beta arises from sampling: if the data are little informative U_beta is large and the method is self-critical in this respect as we can assess its role.
The drift uncertainty, U_x, is a part of the collocation error. From the measurement point of view, U_x inform us on how good could be x to calibrate y, this is why we termed it "reducible environmental error".
The so-called irreducible environmental error is the part of the total uncertainty that our model is not able to put in the reducible part. If the model is adequate, that is if the major physical/environmental factors are available for the analysis, may be related to turbulence. Further, one could think of extending the analysis and dividing many additional error contributions terms (turbulence being one). But, our focus and the radiosonde measurements discussed are in more larger/climatic implications and did not necessitate, we believe, the division to turbulent scales.

9.1) Eq. 7. Why is this equation like this?

This is a well known result on the variance of linear combinations. See Appendix 1, last line of pag. 7520.

9.2) What is the "variance covariance matrix of x(.)"?

In functional data analysis the variance-covariance matrix function of x(h) has elements given the covariances among x_i(h) and x_j(h), i,j=1,…,k, for each fixed h, see also Ramsay and Silverman (2005).
10) In the equation U_omega = E(sigmaˆ2(.|x)), what is meant by the expected value of a variance?

Note that sigmaˆ2(h|x) is a variance conditional on the particular values of the environmental covariates x(h), hence it can be averaged wrt to the environmental covariates x(h).

11) S(h) is introduced. Is this used later anywhere in the paper?

It is not used elsewhere in the manuscript but we wrote its exact formula because we wanted to stress the difference between S(h) and U(h).
We will add some comments on vertical resolution as in Q.4.1 above.

12) Section 4. Delta mu, which should be the difference of the smooth profile is now a co-location drift, why?

$\Delta\mu$ is the difference between two smooth atmospheric profiles, μ (h) and μ °(h), is then smooth. The observed collocation drift is $\Delta\mu + \Delta\varepsilon$ which is not necessarily smooth.

13) Section 5, example. Eq. 10. After all the mathematics developed before, the difference in relative humidity between two atmospheric measurements is a linear combination of the absolute profile and the differences of other measured parameters. Was all the mathematics developed before necessary to arrive to this solution?

The simplicity of Eq. 10 is the beauty of the functional approach which considers an entire profile as a single object. In order to properly understand it, one has to consider also the inverse problem, that is the estimation of the functional coefficients β's depicted in Fig.5 and the underlying maximum likelihood estimation algorithm, which solve an optimization problem to compute simultaneously the β's and the γ\'s of Eq.11 and Figg.6 and 7.
Moreover the mathematical approach allows to arrive to the focal point of the manuscript which is the uncertainty budget both in form of profiles (Fig.8) and summary table (Tab.1).

14) Eq. 10. The differences in relative humidity between two instruments are a linear combination of the differences of the mixing ratios between these two instruments.
Relative humidity _is_ directly calculated from mixing ratios, so a statistical relationship with a high correlation is trivially expected.

Relative humidity is defined as the ratio between the mixing ratio and the saturation mixing ratio: W/Ws.
This means that the differences in relative humidity between two instruments are a linear combination of the differences of the mixing ratios between these two instruments if and only if the saturation vapor pressure for the two instruments is the same.
Ws is a non-linear highly variable function of temperature. Two sites separated by 52 km have two different temperature profiles and therefore two different Ws profiles.
Obviously, these differences should be larger in the boundary layer than in the free troposphere.

15) Eq. 11 is not clear what is meant with this. And again, what is the final result of the example and what is the final objective of the paper?

An explanation to Eq.11 will be added and the fact that the objective of the manuscript is to introduce a general tool will be better explained.

16) In the conclusions the main objective of the paper should be explained and whether these have been achieved in the example.

Conclusions will be improved.