Atmos. Chem. Phys. Discuss., 12, C175–C183, 2012
www.atmos-chem-phys-discuss.net/12/C175/2012/

Atmospheric Chemistry and Physics Discussions

Interactive comment on "How relevant is the deposition of mercury onto snowpacks? – Part 1: A statistical study on the impact of environmental factors" *by* D. A. Durnford et al.

Anonymous Referee #1

Received and published: 20 February 2012

Principal evaluation criteria:

Scientific significance: Good (2) The paper does not present really new data or concepts, but it attempts to utilize the sum of existing (reliable) data and known processes affecting Hg cycling to establish the relative importance of these processes. It is a necessary step in constructing better predictive models of Hg behaviour in the environment, and the study could therefore make an important contribution on the subject.

Scientific quality: Fair (3) I find several flaws and weaknesses in the statistical analysis. The overall approach adopted may be sound, but the application of the statistical methods which are central to the study lack in rigor, which undermines, in my opinion,

the robustness of the conclusions. See below.

Presentation quality: Good (2) The text is well-written, and easily readable. Likewise the tables are simple, explicit, and easy to read. However I have issues with some of the figures, which to me do not convey all the necessary information that would allow readers to evaluate the authors' contentions. See below.

Specific evaluation criteria:

1. Does the paper address relevant scientific questions within the scope of ACP ? YES 2. Does the paper present novel concepts, ideas, tools, or data ? YES (but only partly) 3. Are substantial conclusions reached ? YES 4. Are the scientific methods and assumptions valid and clearly outlined ? NO 5. Are the results sufficient to support the interpretations and conclusions ? NO 6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results) ? NO (lack of details on statistical algorithm use in MLR analysis) 7. Do the authors give proper credit to related work and clearly indicate their own new / original contributions ? YES (but lacks references on statistical methods) 8. Does the title clearly reflect the contents of the paper ? YES 9. Does the abstract provide a concise and complete summary ? YES 10. Is the overall presentation well structured and clear ? YES 11. Is the language fluent and precise ? YES 12. Are mathematical formulae, symbols, abbreviations, and unit correctly defined and used ? YES (but maths lack details) 13. Should any parts of the paper (text, formula, figures, tables) be clarified, reduced, combined of eliminated ? YES (methods section need elaboration) 14. Are the number and quality of references appropriate ? YES (but lacking references on statistical methods used) 15. Is the amount and quality of supplementary material appropriate ? N/A

Comments:

While I appreciate the intent of the authors, I find that the paper suffers from a lack of rigor in the way the chosen statistical methodology was applied.

First, when computing correlation coefficients between the environmental variables and the mercury variables, none of the bivariate scatter plots are presented, only the mean regression coefficients (Figures 2 and 5). I appreciate that the intent of these figures is to summarize graphically the correlation coefficients. But without showing scatter plots of the predictor-observation pairs, it is impossible to judge if the values of the correlation coefficients given on these figures truly reflect a linear correlation (the requirement for MLR, see below). A classic example of a "false" colinearity occurs when most of the data are closely clustered, except for one outlier. A simple regression of such data will produce a high correlation correlation coefficient, but this does not imply a true colinearity. Testing for colinearity requires the use of standard statistic (example: t-test) and a defined level of significance. None are given here.

Secondly, to obtain meaningful results with MLR analysis, a number of assumptions have to be met, and none of these are verified or tested in the paper. If some or all the assumptions are not satisfied, the results can under/over estimate the strength of relationships between individual predictors and predictands, and lead to erroneous conclusions about the underlying physical relationships.

Critical underlying assumptions in MLR models that should be satisfied are:

1) Assumption of normality: It is assumed that all variables are normally distributed. While this may be reasonable for some variables in the present study (example: average wind speed WdSpAv, surface-level temperature SfcT), it is much less obvious for others, for e.g., the frequency of solid precipitation above a certain amount over a given time period, PrF24h, PrF6h. In the latter case, my expectation is that these frequencies should vary non-randomly as a function of the mean precipitation rate, since high-precip. rates represented the area-integrated tail of the (presumably normal) overall frequency distribution of precipitation. The derived frequency distribution of the high-precipitation episodes is therefore unlikely to obey a normal law. It strikes me also that since all the predictors used here are model-generated (as opposed to being empirical observations), the error distributions about the GRAHM model predictions

should be discussed, so the reader can asses whether the assumption of normality is, or not, satisfied.

2) Assumption of linearity in the relationship between the independent variables (predictors) and dependent variables (predictand): Here the authors have "filtered", to some extent, their choice of variables to use in the MLR model by excluding those that were only weakly linearly correlated with the observations (absolute value $\geq$ 0.35, although the choice if this treshold seems rather arbitrary). But in the present case, the assumption of linearity is almost certainly violated for at least some of the predictor-predictand relationships. For example, the fractional loss of total Hg from surface snow is probably strongly affected, in a non-linear way, by the strength of turbulent GEM diffusion through snow driven by surface wind speed WdSpAv (wind-pumping effect; see for e.g. Lia and Tan, 2008, ACP 8: 7087-7099). Also, if Hg deposition results from a combination of both wet and dry deposition processes, then the mean concentration of total Hg in the seasonal snowpack will almost certainly vary non-linearly with total precipitation (PrTot) and/or snowpack depth (SnoDp), because the concentration of Hg in falling snow (wet deposition) is determined by nucleation, adsorption or scavenging processes, which are not linearly related to the snowfall rate.

3) MLR models also assume that the errors on the predictands (observations) are uncorrelated with the predictors (variables). This is usually checked with a residual analysis, but this was not done here. If the distribution of residuals from the regression of an observational variable with a given predictor are found to be non-random, the predictor can be transformed to make it conform with the MLR assumptions (e.g., by "studentisation").

4) Next there is the assumption of homoscedasticity, i.e. that the variance of errors in the variables is constant, or the magnitude of errors is independent of the magnitude of the actual values of the observations. Again, this can easily be tested by scatter plots and an analysis of residual distributions. Based on my own (admittedly rather limited) experience with THg in snow, I believe these data are heteroscedastic, i.e. the variance

of the data increases considerably with higher THg values, which, in the context of a MLR analysis, would possibly justify performing a log-transform of these data prior to analysis. This should be verified for THg and other variables before applying MLR analysis.

5) The predictors should be independent. This is clearly not the case for many of the variables listed in Table 7. For example, if the pool of atmospheric Hg is limited at any given time, the respective amounts deposited by wet and dry processes (DoxDp and WOxDp) will be inter-dependent: The more is deposited by wet processes, the less will be available for dry deposition, and vice versa. The amount of short-wave radiation absorbed at the snow surface (SW) is also by definition dependent on the surface albedo (Ab). The average wind speed (WdSpAv) is obviously influenced by the frequency of high-speed winds (WdSpF6). Likewise the mean snowpack density (SnoDn) is influenced by the snowpack depth (SnoDp), and the latter is also dependent upon total solid precipitation (PrTot). The inter-dependence of some of these variables can generate multicolinearity in the predictors, which weakens the significance of the MLR results in the sense that while the analysis may still produce a good model fit to the ensemble of predictors (explaining most of the variance in the observations), the relative apportionment of the variance between the predictors may be meaningless, and this would defeat the purpose of identifying those which have the strongest control on the process being examined. The possible redundancy of some predictors is acknowledged by the authors (page 14, page 27) but that doesn't resolve the issue.

The considerations above lead me to think that in this study, many or all of the important assumptions that validate the use of MLR analysis are violated. At the very least, more information needs to be presented to justify the use of MLR. I also wonder why none of the MLR model regression coefficients are reported (for e.g., in a table), as these should also give some indication of the relative strength (relevance) of the various predictor-predictand relationships. These coefficients should be presented, and their significance should be tested (for e.g., using a t-test). Moreover, it would be appropriate

to explain what algorithm used in performing the MLR analysis, more specifically, which is the parameter that is being minimized ?

Lsttly, there is the important issue of model validation: A MLR model that predicts observations of a variable (predictand) with great fidelity implies that the latter can be adequately described using a combination of linear relationships between the predictand and the chosen set of predictors. However, the performance of the model should be tested against observations that were NOT included in the "training set" used to generate the models' regression coefficients. Otherwise we have a circularity. This is a common problem when using MLR for time series analysis, and there are various ways to address it. One can, for example, leave out a randomly-selected subset of observations for a given predictand when performing the regression, and then test the model performance in predicting those observations only. The process can be repeated by excluding, in turn, any subset of the observations. One can also generate random sets of observations and test how well the model performs against such "dummy" observations, then calculate what is the likelihood (probability) that the apparent correlation between model results and real observations may be due to sheer luck, which is the Monte Carlo validation approach. None of this was done here. The MLR model was "trained" to reproduce observations, but its performance was only tested against the observations with which it was trained. This is not a proper validation. In the present case, for at least some of the mercury variables (e.g., 24-hour losses of Hg from the snowpack) there are many more predictor variables than there are observations. This can lead to over-fitting using MLR. In such a case, partial least-square regression might be a more appropriate approach. In any case, some model validation should be performed.

These shortcomings are important because the MLR analysis is a central element of the study, and a large part of the discussion rests on its validity. If the key assumptions above can not be verified, or are violated, an alternative choice of method for the multivariate analysis may be warranted, for example ridge regression (for correlated predictors) or regression on principal components of the predictors, instead of individual predictors. There are plenty of scholarly texts available that provide guidance on so-called "regression diagnostics" that can assist in selecting the right method. See for example Belsley et al. 1980. Regression Diagnostics. New York: John Wiley & Sons. Many of the recipes associated with these diagnostic tests are now routinely integrated into statistical software packages such as SPSS.

On another topic, I doubt if there is any real value in including the results of the MLR analysis for both Set1 and Set2 observations. At least some of the observations of Set1 are almost certainly unreliable, like the 1970s data on THg in long-term cryospheric records. This is in fact acknowledged by the authors. But the inclusion of these results in the analysis of Set1 observations almost certainly biases the MLR-inferred relationships not only for this particular type of observations, but for others, since the success of the model to replicate any of the predictands is partly controlled by the strength of the linearity in all predictor-predictand relationships included in the model (this is intrinsic to the MLR procedure). Furthermore, at least part of the discussion of Set2 results (pages 20-21) seems to simply re-iterate the reasons that were offered beforehand for excluding particular observations from the dataset when preparing Set2 calculations. It doesn't seem to bring anything really new to the story as told.

The one seemingly important conclusion that the authors draw from their comparison of Set1 and Set2 calculation and results, is that removal of observational data from sites where halogens may have promoted the retention of Hg in snow improves the performance of the MLR model in predicting observations. The authors use this result to stress the importance of halogens as a controlling agent for mercury levels in snow (page 31). This may indeed be tgrue, but the inference that is made here contradicts one premise of the stated approach, which assumes that neglecting an important environmental variable in the MLR model should lead to poorer predictions by the model, not improve them (as stated in section 2.3). In fact, I suspect that removing certain types of observations (e.g., THg in snow) for specific geographical sites, while keeping

the other observations for these same sites (as was done here), introduces a bias in the MLR model performance that should be avoided. Possibly a more correct, rigorous and informative way of testing the relative importance of the different environmental variables included in the MLR analysis would be to perform an iterative calculation in which the performance of the model to reproduce any of the 5 key observation types is re-evaluated after successively adding (or removing) a predictor variable from the model.

There are a few other points in the text on which I am in disagreement with the authors:

On page 21, the authors state that the sub-horizontal alignment of data points on Figure 3b (pertaining to concentration of THg in surface snow) suggests a latitudinal characteristic. That inference is far from obvious to me. In fact, my reading of this figure is that the MLR model produces similar results for THg in surface snow regardless of the latitude of the observations. Only one data point from Antarctica departs from this general trend. I fail to see how this would lead the authors to infer some latitudinal relationship.

On page 27, the authors state that "...mercury concentration (in seasonal snowpacks) tends to increase with latitude". But this is far from obvious when one looks at Figure 4 of Dunford and Dastoor (2011), which show the data used in the present analysis. What may be true is that certain characteristics of high-latitude environments (such as the high levels of halogens in snow on or near sea-ice) lead to higher retention of Hg in seasonal snowpacks, but this is not a generalized latitudinal relationship. In fact, looking at Figure 4 suggests that overall, THg in seasonal snowpack in the Northern Hemisphere is poorly correlated with latitude. And with only two data points in the Southern Hemisphere, nothing can be said conclusively.

In view of the comments above, I recommend that the authors expand their discussion of the statistical methodology used by addressing the assumptions that underlie the methods, and evaluating the robustness of their correlation and MLR analysis results. I don't think the approach used here is necessarily flawed, but it definitely lacks the appropriate rigor. A great deal of effort went into interpreting (and quite possibly over-interpreting), in physical terms, the results of a statistical analysis that, in my opinion, was done in a somewhat sloppy fashion. More rigor in the method might go a long way in sorting out what is truly relevant (and what is statistically not meaningful) in the ensemble of possible relationships between the variables.

Interactive comment on Atmos. Chem. Phys. Discuss., 12, 387, 2012.