**Atmospheric
Chemistry
and Physics
Discussions**

Interactive
Comment

Interactive comment on "On the relationship between total ozone and atmospheric dynamics and chemistry at mid-latitudes – Part 1: Statistical models and spatial fingerprints of atmospheric dynamics and chemistry" *by* L. Frossard et al.

Anonymous Referee #2

Received and published: 12 July 2012

The discussion paper by Frossard et al. presents statistical methods to detect anomalies in ozone time-series and applies these to NIWA total ozone data. The scope of the paper fits well in ACP, and the contents seem to be correct, but the style of presentation does not really address to the majority of the readership of ACP. There is some disbalance between the level of explanations of the statistical contents and that of the atmospheric science content: While the statistics section is extremely demanding, with a lot of unexplained technical terms assumed to be known which might be clear for statisticians but not necessarily for atmospheric scientists, and a lot of details which

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

make it difficult to recognize the greater whole, in Sections 4.2.x several atmospheric phenomena are explained with much less details assumed to be known. I like these summaries of atmospheric phenomena, but in a journal on atmospheric chemistry and physics I would expect rather the opposite: Less statistics but instead atmospheric phenomena assumed to be known. At many places the reader might have trouble "to see the wood for the trees" because he might be overloaded with details but is not really guided through the detail aspects and equations. In many places there are cascades of definitions which make the paper hard to read in sequence and force the reader to browse back and forth. Further, the paper is hard to read because of frequent very long sentences. Since I consider the content of the paper as important and useful, I recommend publication of this paper after reworking for readability and consideration of the following specific comments:

p 13163 l1: The term "model" has a different meaning in statistics than in atmospheric sciences. Models in atmospheric sciences are mostly deterministic, not probabilistic. Thus the use of this term is misleading. I suggest to specify the models already here to avoid misunderstanding. At least say "statistical models".

p 13163 Abstract: The abstract should summarize not only the method but also the results.

p 13164 l29: I suggest comma after particular.

p 13164 bottom - p 13165 top: long sentence.

p 13165 l 18: the acronym ARMA should be defined. It has been defined in the abstract but not in the body of the paper. The body of the paper should be self-containing and not rely on the abstract. Further, at the end of the introduction it would be advantageous to make a clear and simple statement which question shall be answered by which method.

p13166 ll22: I suggest comma after study.

p131167 l1-4: This short motivation of the choice of methods would be useful already at the end of the introduction. Otherwise the reader might be at a loss with the technical terms at the end of the intro. Here the definition of the acronym ARMA is given, but the acronym is already used in l18 of p 13165.

p 13167 l6: Please make a short statement here what the purpose of r-largest order statistics is. E.g.: The r-largest order statistics treats the probability that the r largest members of a sample are larger then a threshold ... or something like that. Such an introductory statement will help the reader to better understand the details listed below. The current text immediately starts to discuss detail aspects without any more general introduction to this issue.

p13169 l4: is the lefthand expression of Eq(1) the log likelihood? If so, then please insert this expression after the term "log likelihood" in the text.

same place: "log likelihood" appears to be slang to me. Shouldn't it read "logarithm of the likelihood". Or if the technical term is meant, use hyphen: "log-likelihood".

l1369 Eq.1: There is a cascade of definitions, which extends down to line 8 on page 13170, which all are necessary to understand Eq. 1. Couldn't this part be reorganized from a top-down style to a bottom-up style to avoid that the reader encounters so many yet undefined quantities? I.e. go from Eq (3) to Eq (2) to Eq (1) and then to the z-test?

13169 12ff: "normality" can be anything. Probably it refers to normal distribution (and not normality in the sense of orthogonality or in the sense of normalization). What does the MAXIMUM likelihood estimator mean here? "The asymptotic normality of... which should be applicable... allows us." This is certainly all correct but if you say: The large number of months (b=348) justifies, according to the central limit theorem, to assume that ... follow approximately a normal distribution. Thus the significance of the parameter estimates can be tested with a z-test. (To me it is not quite clear here what "the estimate" is. Can this be specified?); By the way: why is the central limit theorem applicable? Please specify which are the additive random processes. The sum in Eq

1?

13169 12-16: This paragraph assumes that the reader knows all details on the z-test. While the z-test indeed is quite a standard tool, it is often used without referring to its name, and many people who use it might not know the term z-test. Thus some explanation might be useful. Please state explicitly that the p-value is the probability that the estimate (whatever this is) divided by its standard error is consistent with the null hypothesis (or any other similar statement in the case that my suggestion is incorrect).

13169, l15: Calculation of the standard error may be difficult. Is there reason to assume that the elements of the sample are independent?

13169, l20: This would all be clearer if it was stated somewhere what the order statistic predicts: The probability that the r-largest values exceed a certain threshold, or the most probable value of the extreme value, or anything esle? What does the $\sim$ Symbol mean here? Have y(x,t) and the GEV the same units?

13169 l2: What is the nature of μ? Is it an ozone column? An extreme value? or a probability?

13169 Eq 2 and below: Is the design matrix Z(t) really necessary? To me it seems to add a lot of unnecessary formalism and further technical terms which are not helpful for readability of the paper. I do not see where the design matrix is further used in the paper, since Eq. 3 provides directly the μ needed in Eq 1.

13170 l5: "month" is ambiguous because it is not clear if it is distinguished between the same month in different years. "each of the 348 months" removes this ambiguity.

13170 l14: "best": Has a certain norm been minimized or is "best" a subjective choice?

13170 Eq3 and above: This is called "model" and this is in line with the typical use of the term "model" in atmospheric sciences. I would call it "location-dependent model of ozone variation with time" or something like that. The use of the term "model", however, seems to clash with the term "model" elsewhere in the paper.

Interactive
Comment

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

13170 Eq3 I do not see the x-dependence in this equation. It is mentioned later in the text that the beta-values are x-dependent but it would be clearer to see the x-dependence already in the equation.

13170 Eq 3: please specify what Eq 3 predicts. After a while I figured out that it simply predicts ozone column amounts but this should be explicitly stated. According to the text before one might also expect that it predicts extreme values, probabilities or anything else.

13171 l2: The statement that the betas are time-independent could be omitted if the dependences of beta are explicitly stated in Eq 3.

13171 l6 "may be nested" sounds a bit vague. Please make an explicit statement that this is actually done for certain purposes and avoid the "may be" statement.

13171 l25 "models were also fitted" not clear what is fitted here. The beta coefficients of Eq 3? Please be more specific.

13171 l26 "standard errors" of what? (With a more instructive description of the method, particularly the z-test, this question may become obsolete).

13172 l1-9 This text suggests that multiple linear regression models never consider autocorrelations. This is, however, not true, there exist multiple linear regression models which support the use of the full covariance matrix of the data errors (c.f. von Clarmann et al., Atmos. Chem. Phys., 10, 6737-6747, 2010) and there exist methods to determine the covariance matrices from the data (e.g. AR(1) method as used by Vyushin et al., JGR 112, D14307, 2007). Are linear regression models considering covariances not also called multiple linear regression models?

13172 l10ff: please describe shortly the main characteristics of ARME methods. What is their advantage compared to other methods? Is it that they detect and support time-dependent correlations?

13172 Eq 4: What is the nature of Y? How is it related to y? Why is it printed as a

capital letter? Is it a scalar of a vector? Does this (y vs. Y) follow some notational convention about distribution and their realizations?

13172 Eq 5: Is the nature of $Z(t)\beta(x)$ the same as in Eq. 2? If not, please specify the differences.

13172 Eq 6: What is the meaning of this term? Probably it describes the autocorrelated noise but some explanation would be helpful.

13172 l21: What does the $\sim$ with the superscript "ind" mean?

l13172 l1: Here it is stated that the term $Z(t)\beta(x)$ is the same as that used in the order statistics Section. Does that mean that it predicts a kind of time and location corrected expected value of the same dimension of y which is then subtracted from the measured value y to have zero-expectation? Eq 1 suggests this. Now, in retrospective, pages 13169 and 13170 make more sense to me, but it would be most helpful to state already near p 13168 in easy words what the general rationale of the procedure is. Now I understand you fit coefficients beta to a regression model and you apply the r-largest order statistic to the residual between the regression model and the measured data. Is this correct? If so, then please state this in easy words early in the paper to guide the reader through the equations. It took me multiple reading to get this - and I am still not sure if I got this correctly.

General question: Are the autocorrelations in the residuals only an issue in the analysis of mean values and not in the analysis of extremes?

Which problem is solved by the ARMA approach? To determine the covariances of the residual, or to consider these in the fit, or both?

13173 l 10: I doubt that the Akaike information criterion is known by the majority of the readers of ACP. Please explain.

13174 l 10: The Bonferroni correction might not be known by the majortity of readers of ACP. Please explain. What are the underlying assumptions?

13174 l15: Please guide the reader through the results by first stating what the purpose of the analysis is, what hypothesis is to be corroborated/falsified, which relation is to be demonstrated. Otherwise the reader might get lost in the quantitative details and lose the thread.

Sect 4.2. ff: Although the patterns discussed are widely known, I like that their main features are summarized here, because this makes the paper better accessible. I wish the statistics section would be written in a similar way.

13182 l4: Is ozone really transported that far? According to Brasseur and Solomon, Aeronomy of the middle atmosphere, Springer, Sect 5.2.3., this view could be somewhat over-simplified.

Figures are generally too small. Axes labels are hardly readable.

Figures 2 and 3: panels b and e should be discussed i more detail. I wonder that correlations appear not even in tendency to be autoregressive. In fig. 2 there is a maximum at time lag 12 months. Does this suggest that there is a problem with the annual cycle?

In summary: Although I am interested in statistical methods and support the publication of their use for atmospheric sciences, I had a very hard time to review the paper. I suspect that many readers of ACP might have similar problems and I strongly recommend to present the contents of the paper in an easier accessible way. Perhaps it would help if those coauthors who are not that deep in statistics but more on the application side checked the final version of the paper for clarity and readability. After quite a while I understood the following (I still may have misunderstood something): 1. A parametric model of time dependence of ozone column amounts is built. Location-dependent coefficients $beta_i$ are fitted to the NIWA data. (Eq. 3) 2. The residual between the modeled and measured ozone column amounts is calculated ($y-\mu$). 3. The residual is analyzed both in terms of extreme values (Sect. 3.1) and mean values (Sect. 3.2) 4. The probability that the 3 largest (smallest) ozone values in each month are consistent

with natural statistical variability and the parametric model is assessed (Eq 3 and subsequent z-test). Fingerprints with significant deviations are identified... 5. Similarly for mean values. ... Is this the guideline through the paper? If so, why not describing it like this? This way to organize the contents requires much less cascades of definitions and would be, to my judgement, much more intuitive for the atmospheric scientist.

Interactive comment on Atmos. Chem. Phys. Discuss., 12, 13161, 2012.