Rebuttal On RIVM's Critique Of Briggs et al. AERIUS/OPS Model Performance

William M. Briggs
matt@wmbriggs.com
Michigan

Jaap C. Hanekamp
j.hanekamp@ucr.nl; hjaap@xs4all.nl
University College Roosevelt, Middelburg, the Netherlands
Environmental Health Sciences
University of Massachusetts, Amherst, MA, USA

Geesje Rotgers
grotgers@gmail.com Stichting Agrifacts (STAF)

October 5, 2022

Abstract

Keywords: AERIUS/OPS, Atmospheric nitrogen, Model verification, RIVM, Skill

1 Summary

The RIVM responded to our paper"Criticizing AERIUS/OPS model performance" [1]. We rebut those critiques below.

In brief: RIVM agrees that OPS performs poorly in the short term, but believes it performs well over long periods. Yet they failed to address our demonstration on how averaging over these longer terms can cause spurious "good results." That is, this averaging causes OPS to produce bogus results.

Alarmingly, RIVM does not understand skill: we prove a simple mean model beats OPS often. This simple model takes the observed averages as predictions. RIVM says these averages are not available, and so this simple model would be unavailable for policy, but that is false. Those averages are right there in OPS, as we show in model run of OPS. And we show here that even in times and places where averages are not available, good guesses of them are adequate.

It is still true that for the OPS model runs we made, one time with 400 cows at a farm, another with half that, and another still with one quarter, the nitrogen deposition (per ha, per year) shows trivial, unmeasurable differences.

That being said, uncertainty in OPS model is not being sufficiently considered, and is substantial, aggravating previous observations. Taking proper stock of uncertainty is crucial because this model is being used to make far-reaching policy decisions.

The solution is to design experiments and independently test OPS on completely new data. Pending those experiments, OPS must be shelved as to forestall extensive societal and economic damages resulting from its continued use in policy making.

2 Introduction

After we published on Research Gate the article "Criticizing AERIUS/OPS model performance" [1], RIVM submitted a critique, or reaction, to our paper on their site. [4]. This is our rebuttal to that reaction.

The Reaction was given in Dutch. To make our rebuttal more widely accessible, we translate quotations from the Reaction into English.

The numbers answering comments are ours, and do not appear in RIVM's reaction. They mostly use headers and bullet points for specific criticisms; however, numbers make for easy organization. These number all important reactions.

1. Introducing our paper, RIVM says, "This is not an article in a scientific journal." Neither is their Reaction. But I am sure they, and we, agree that we would never engage in the appeal to authority fallacy.

2. RIVM says previous studies of AERIUS/OPS (or OPS for short) performance were done using "generally recognized and regular scientific criteria". We agree. And we showed how and why these are inadequate.

3. In our paper, we discussed a simple mean model. This greatly confused RIVM. But it is simplicity itself. The RIVM also says the mean model "has no predictive value." This is false.

The mean model is just the mean, or a guess of a mean, of what future observations will be. It is used as a forecast or prediction of future observations, in the precise same way as the OPS, or any complex model, is used. The simple mean model is a constant prediction that all future observations will equal previously observed mean.

Mean models are common and useful. For instance, we don't know what the exact temperature will be, but it's a good bet October will be colder than September in the Netherlands. Right now, writing in August of 2022, using that model, we predict October will be colder than September. Policy can be made on that model; and, indeed, is. Especially by farmers.

The mean is possibly not as accurate as a complex meteorological model based on thermodynamics, but it (the mean) is still an excellent guess in the absence of that complex model.

We can, and should, compare a mean model with the more complex model. If it turns out, using the very measures advocated by RIVM, the mean model *beats* the complex model, we should not use the complex model. We would do better with the mean model. This is obvious. We say the complex model does not have *skill* with respect to the mean model.

Skill is a common in weather, climate and atmospheric deposition models. We supply many references in our paper to prove this. We could supply many more, or RIVM can look them up on their own.

Skill is easy to understand. It is a silly argument to say that because it has not been used before by RIVM, it should not be used now, or that it is "not relevant".

This is a dodge, a way to escape the consequence of the poor performance of the OPS model.

Below, we show how skill works in detail using the mean model, and using data supplied by RIVM themselves. Skill as a concept is exceedingly useful; indeed, necessary.

4. RIVM says, "While Briggs et al. seem to believe that international models comparable to OPS are 'all just as bad', the RIVM sees, on the basis of the evaluations performed and key figures, that the models are all (with

their own strengths and weaknesses) useful in calculating air quality and deposition."

This is true. Many interesting things can be gained from models of all kinds. However, that doesn't imply any of these models should be trusted to make predictions, especially considering policy implications. Any model needs to prove itself in making skillful predictions of observations never before seen or used in any way.

RIVM says they were, and are, "a strong supporter of this" testing. This is fine. We would be delighted to help design a set of experiments to test OPS that could be evaluated by neutral third party judges.

5. RIVM concludes their general comments by saying that they have "not found any arguments in the study by Briggs et al. for adapting its working method and the current and planned model developments to OPS."

We agree that they have not. But it will be our pleasure to explain our methods to show what they missed, which is most of our comments and critiques.

We do so next answering specific comments.

3 Rebuttals To Specific Comments

We here reply to all important specific comments.

6. RIVM agrees that OPS performs poorly for short-term forecasts. We agree. And they say OPS works better over long-term forecasts, such as monthly or annual forecasts. This may well be so, but it is not proved.

We showed how averaging can lead to spurious improvements. RIVM did not respond to our demonstration of how simple it is to generate spurious correlations by averaging noise, and how this averaging makes models look better, but only cosmetically. The RIVM has not shown that OPS can make independent long-term predictions. This we can discover only through the kind of independent planned experiments we suggest.

RIVM says, "regular calculations of air quality and deposition are based on annual average values. The use of average values in the tests and validations is then no problem." Exactly so.

Even better, these average values *are* the "mean models" we speak of. So it is nice to see RIVM can make use of the means if need be.

7. There is a question of which validation measures to use. RIVM uses those noted in our paper (which we do not repeat here). These measures

are used by others, too, as they and we agree. But this does not make them good measures: they are poor validation measures for the reasons we demonstrated.

They do little more than measure differences in averages of the model and observations. They cannot capture functions of the size of error, or the dependence of error on the magnitude of the observations. Some of them are not proper, in the mathematical sense, as we argued. They cannot capture departures from calibration or precision, again as we argued.

Here we introduce a more robust measure, one common in the verification literature, the complete rank probability score (CRPS); see e.g. [3, 2]. Let x be the model prediction, possibly given in probabilistic form, with cumultatve distribution function (CDF) F_x, and let y be the observable (a single number), then the CRPS is:

$$\mathrm{CRPS}(F_x, y) = \int_{-\infty}^{\infty} \Big(F_x(x) - \mathrm{I}(x - y) \Big)^2 dx, \tag{1}$$

where I() is the indicator function (equalling 1 when its argument is true, else equalling 0). If the model's prediction are points, implying model overcertainty, then F_x is a simple step function, equalling 0 until x and then equalling 1 for all points larger than x. I.e. $\Pr(X < x|\mathrm{model}) = 0$ and $\Pr(X \geq x|\mathrm{model}) = 0$. Then the CRPS becomes

$$\mathrm{CRPS}(x, y) = |x - y|, \tag{2}$$

also know as the mean absolute error.

There are many nice properties of the CRPS given in the references above, though many details are mathematically complicated. It is, however, also very intuitive. It measures the "distance" between the CDF of the model prediction, with the CDF of the eventual observation (always a step function).

Another weakness of the OPS model we did not explore was that its predictions emerge as "points", i.e. single numbers. These, in essence, say "There is a 100% the future observation will be x." Nobody believes this, of course. Information in uncertainty of the prediction is lost when points instead of probabilities are used. It is imperative that this is addressed by the OPS developers

8. RIVM expresses confusion in the simple mean model, as mentioned above. They say the mean model is an "unworkable way of estimating a

value in the absence of measurements...and is therefore unsuitable for policy applications."

This is false.

Above, we gave a simple example of change in seasons as an excellent informal implementation of this model. RIMV's criticism that the mean might not be known with absolute certainty in advance is also weak, as we demonstrate in a moment with the Kincaid data, helpfully supplied by RIVM. The mean is also already available in OPS itself, as we show later.

We give this demonstration using the concept of skill using the CRPS score (though many scores, including the ones RIVM favors, would show the same thing).

Fig 1 is the original Kincaid verification data, taken from the original Kincaid document (and referenced in our paper). It shows several model runs over-plotted on each other. But there is another, more informative way to look at this data (we thank RIVM for providing this data after our paper was out). We show this in Fig. 2.

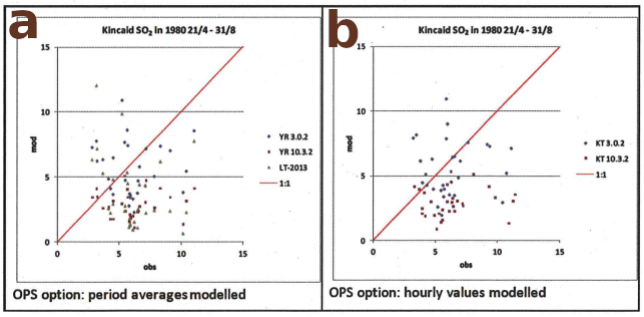

Figure 1: OPS model predictions of SO_2 (y-axis) against observations (x-axis), with dates indicated, for various periods and experimental runs. A red one-to-one line is over-plotted. Perfect predictions would fall on this line. The (a) are period averages, and (b) are hourly values.

Fig. 2 is the same data as in Fig. 2, but shown as individual plots. It is easier to see here that some model runs have substantial bias. This bias was difficult to see when all model runs were plotted over one another.

The average CRPS scores for the OPS and mean model are in Table 1. The mean model here are the means of the observations (i.e. for every observation, the mean model predicts the value will be the mean). As is clear, the mean model easily beats OPS. OPS badly underperforms next to

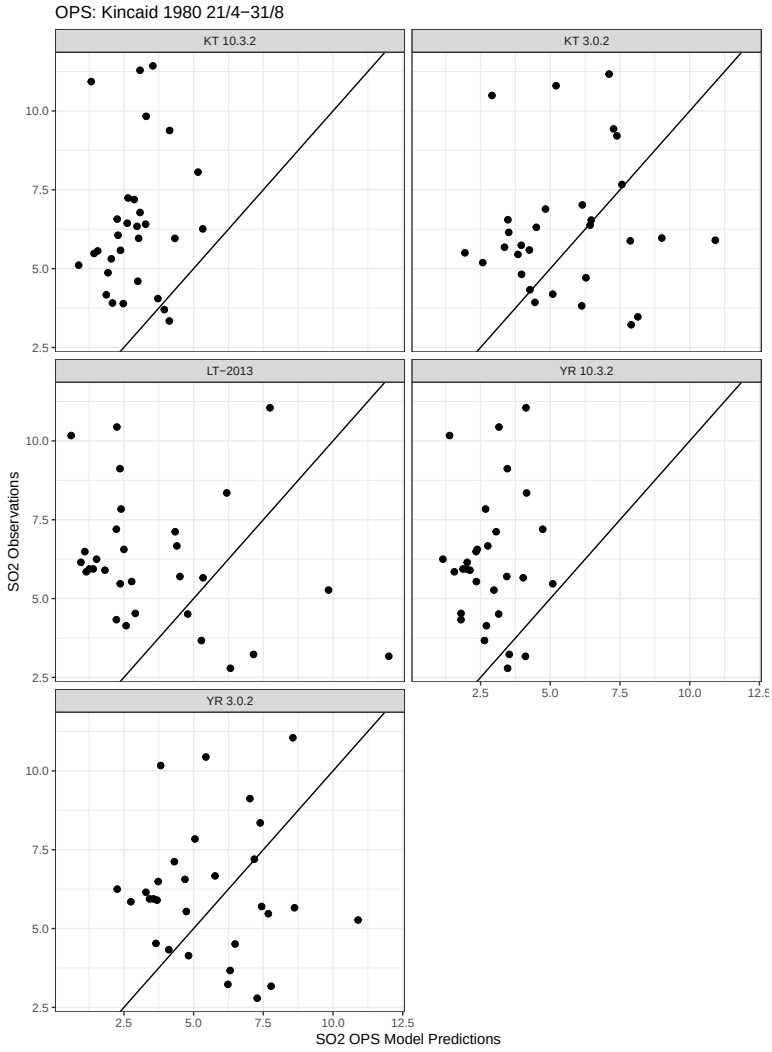

Figure 2: As in Fig 1, but each model run is shown separately.

Table 1: Average CRPS for the Kincaid OPS and "mean" models. Smaller numbers are better. The mean model is far superior to OPS.

	KT 10.3.2	KT 3.0.2	LT-2013	YR 10.3.2	YR 3.0.2
Model	3.57	2.36	3.94	3.43	2.56
Mean	1.66	1.55	1.55	1.55	1.55

it. At this time and location, the OPS model did not have skill compared to the mean model.

The objection the RIVM made was that the mean would not be available in advance, which implies the mean model could not be used. This is not so; indeed, averages are available and already used by RIVM (as we show in a moment).

But suppose the mean were not known with certainty, and a guess instead had to be made of what the mean would be. Then we arrive at Fig. 3.

This shows the CRPS for the Kincaid data, using various assumed means (black line) for the mean model. The OPS model average CRPS is in dashed red, and the observed mean model average is dashed blue. For three of the model runs (KT 10.3.2, LT-2013, and YR 10.3.2) the mean model beats OPS at every assumed mean. The mean model also beats OPS for most assumed means at the other runs.

In other words, we start by assuming we don't know what the exact mean is, but make a guess, and use that guess as the mean model. These guesses (assumed means) become the simple mean model. In all runs, the assumed means have to veer very far off before OPS becomes competitive. This shows that if crude guesses of the future mean used as the mean model are good enough to beat OPS.

Even stronger, to even get OPS started, some kind of notion of area average or mean must be available (see below). So it doesn't appear there are any instances in which the mean model couldn't be computed.

Next we demonstrate the difficulty of comparing models with single scores.

Fig. 4 shows CRPS as the size of the observable (SO_2) for the Kincaid data. Loess smoothers with confidence intervals are also given. It's clear that as SO_2 grows larger, the average OPS error also grows.

Importantly, the error becomes largest at values of SO_2 that are most important. When SO_2 is small, relative model error is not especially important, because SO_2 at small values isn't as interesting. But when SO_2 is large, it becomes crucial for the model to perform well.

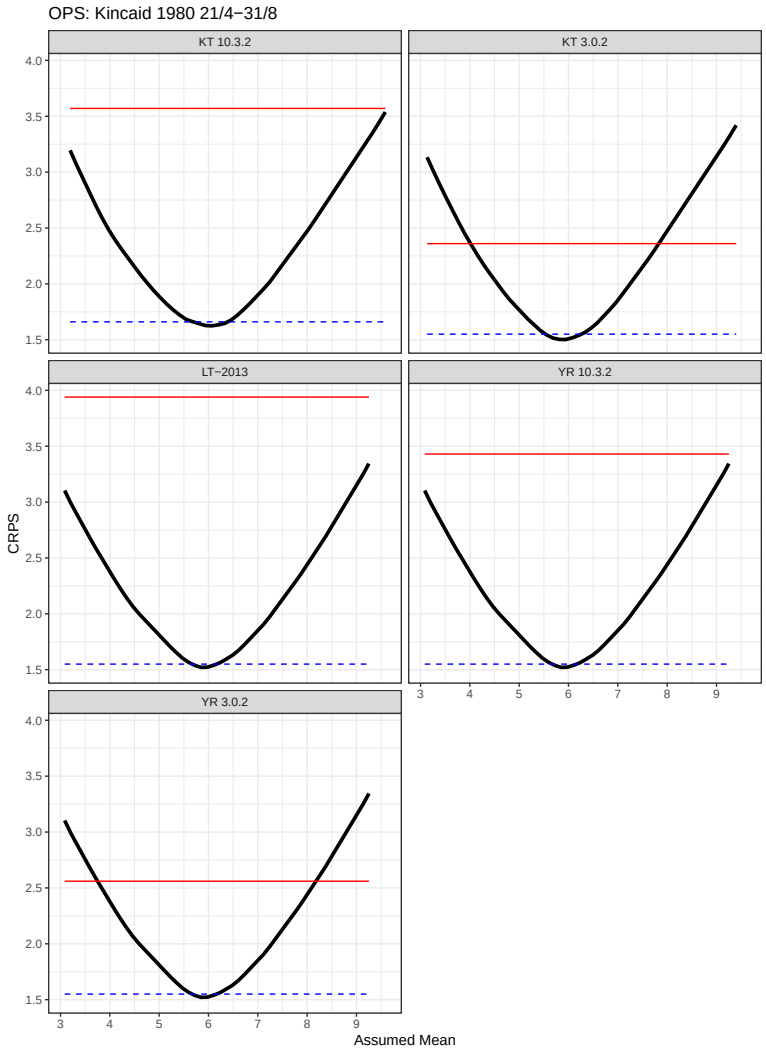

Figure 3: The CRPS for the Kincaid data, using various assumed means (black line) for the mean model. The OPS model average CRPS is in dashed red, and the observed mean model average is dashed blue. Even if the future mean was not known exactly, the simple mean model usually beats OPS.

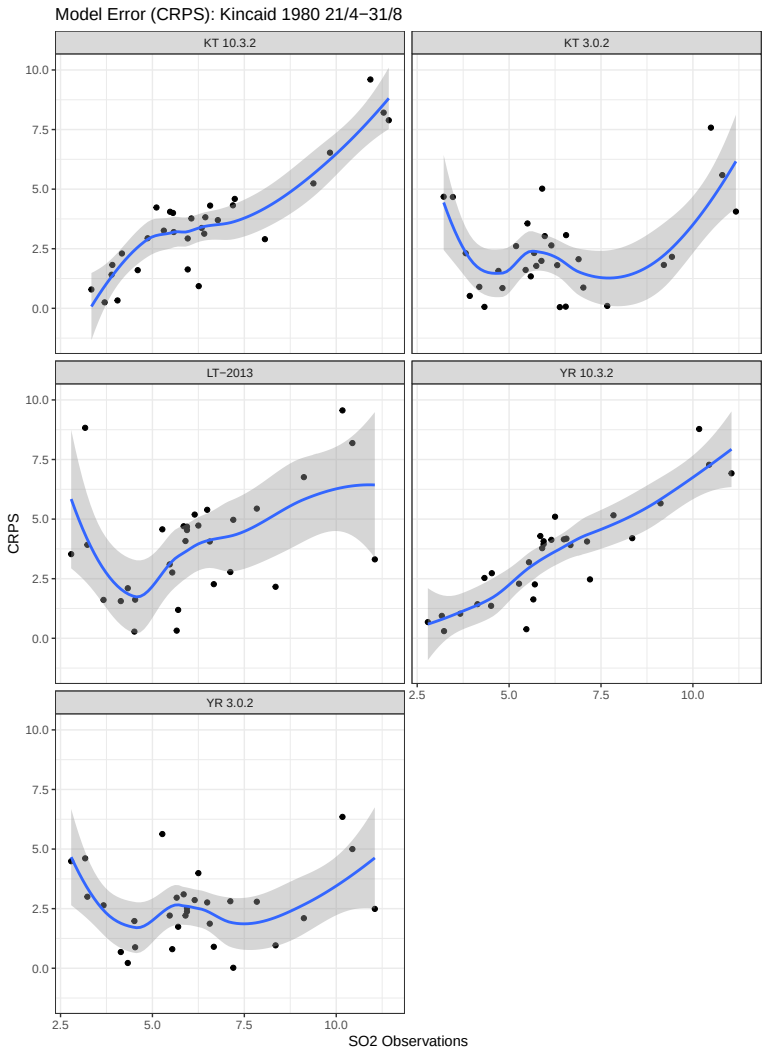

Figure 4: CRPS as the size of the observable for the Kincaid data. Loess smoothers with confidence intervals are also given in blue with gray shading.

We see this error-size dependence in other datasets, too. Fig. 5 is the same as Fig. 4, but for the Falster data. All data sources are described in detail in our original paper.

Once again, OPS model error increases as the observable becomes large, and important. This is even easier to see in the North Carolina portion of the Falster dataset.

Fig. 6 first shows the scatterplot of OPS model predictions and observations. It can be easy to miss, but the predictions are much larger than the observations (note the axis markers).

This is backed up in Fig. 7, which shows the CRPS as a function of observation size.

Again, it is easy to miss the size of the error. The NH_3 measures run from 0 to about 100. But the errors are much larger. Instead of CRPS, we switch to a simple error score of Prediction minus Observation, and repeat the figure in Fig. 8.

The average simple error can be very large relative to the observation, and, of course, increases as the observation does. The OPS model can be off by twice as much, or more, significant errors.

The real problem with this rebuttal, RIVM's critique, and our original paper, is that all these analyses are secondhand. The OPS model runs were largely (or possibly all?) on old, alread known (to the modelers) data. The Kincaid data was from the 1980s, but the models were run decades afterward.

It is too easy to tune models (of any kind), as we described in our original paper, so that the model matches the already observed data. The possibility of confirmation bias, always very real, is with us here.

Again, what are needed, are planned experiments where OPS makes predictions of data never before seen or used in any way. Preferably this is future data, with tests conducted as they are on, say, meteorological models. This is essential to prove the actual skill of OPS to all interested parties.

9. RIVM takes exception with our discussion of a particular run of the OPS model. We repeat the findings here for quick reference:

The first, and perhaps amusing, observation is the "0" cow, or background deposition. Which is, we remind the reader, the mean model. Or a version of one. So RIVM can, and has indeed, provided a mean model—one on which policies are being made.

Now in our original paper, we emphasize that this model run shows a trivial difference in deposition from halving cows, from either 400 to 200, or 200 to 100, or even 100 to 0 or no cows. The deposition from the full 400

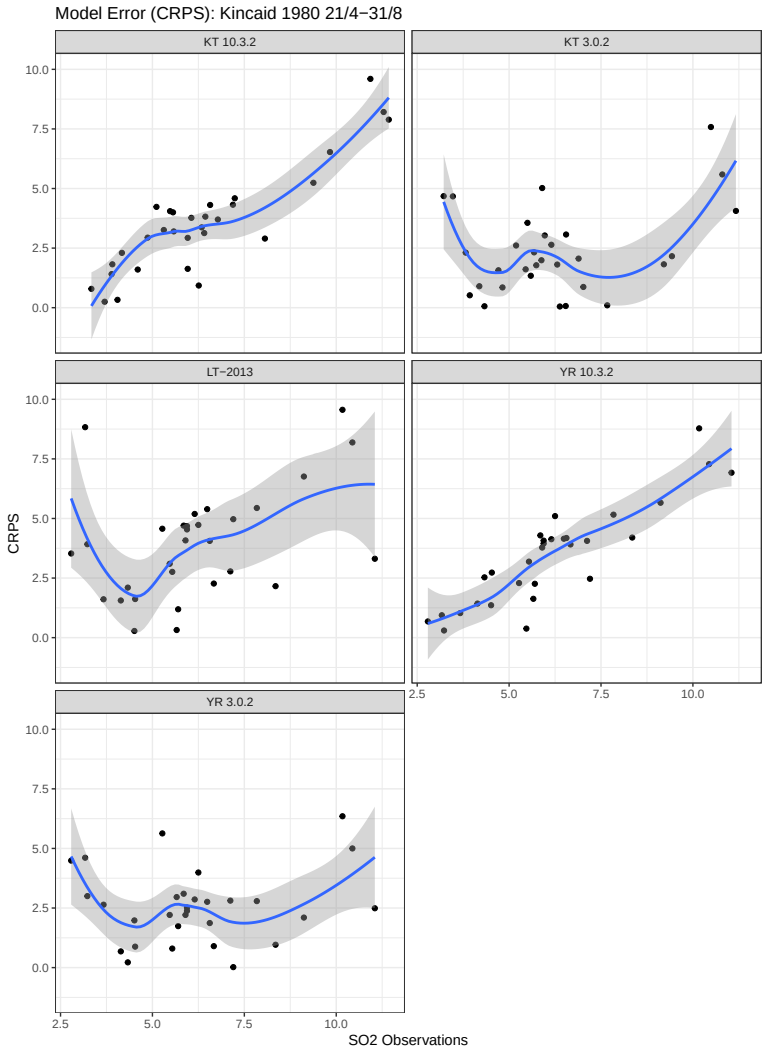

Figure 5: As in Fig. 4, but for the Falster data (NH_3).

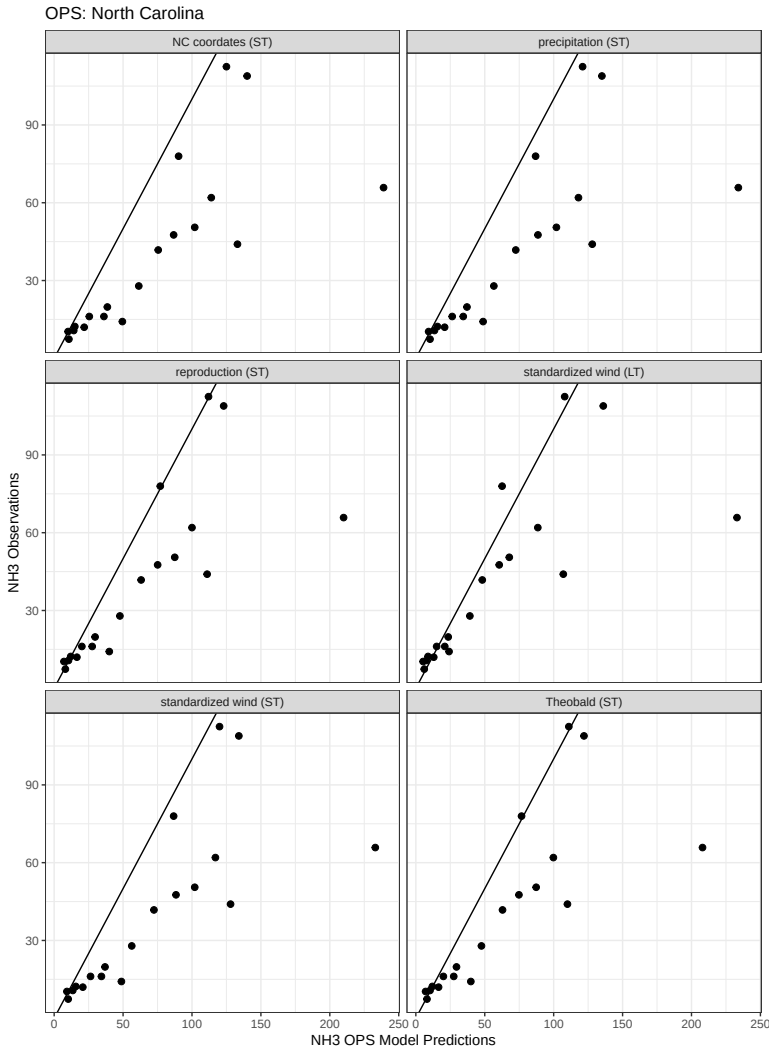

Figure 6: OPS model predictions by observations (NH_3) for the North Carolina data.

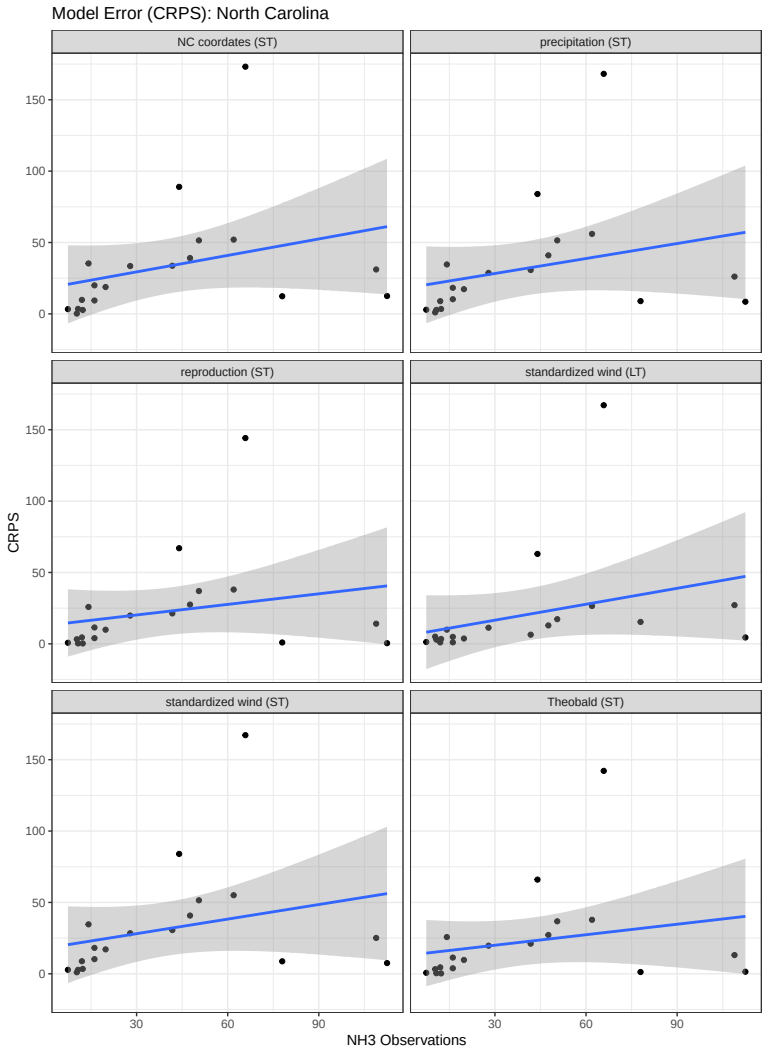

Figure 7: The OPS CRPS error by observation size, as in Fig. 4, but for North Carolina.

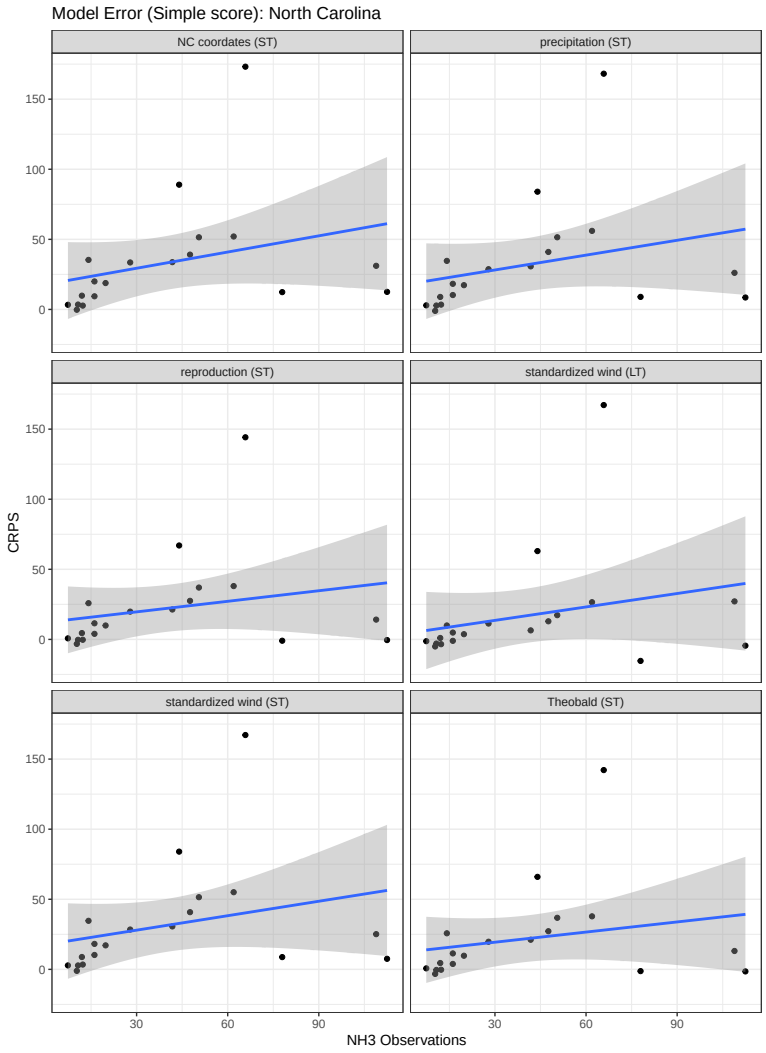

Figure 8: As in Fig. 7, but for the error score of Prediction minus Observation.

Table 2: Results of the AERIUS/OPS model for various scenarios. The "0" cows is the background deposition only. I.e. a simple mean model. Total depositions are in mol N/ha/year, and the estimated deposition just from cows.

Cows	Total deposition (mol N/ha/year)	Deposition from cows
0	1938	0
1	1938.07	0.01
100	1939.52	1.47
200	1940.99	2.93
400	1943.92	5.87

cows is, at maximum, only under 6 mol N/ha/year.

We claim, and it is true, this is a negligible number, and, even if it is perfectly accurate, that it is scarcely measurable to any precision. Could any real field measurement tell the difference between, say, 5.87 and 2.93 mols N/ha/year? No.

That means there is no way to verify the model.

This makes RIVM's comment that we say "nothing about the correctness/defensibility of [the model's prediction] value" as not correct. They also say "Just because a number is small doesn't mean it doesn't exist," which is very true. But irrelevant.

Again, the model may be perfect, but how would we know? We can't measure to the precision required.

10. RIVM says:

> Briggs et al. indicate that models comparable to OPS show the same "performance". In other words, the OPS model does no worse or better than other models that are common in air quality and deposition calculations. While Briggs et al. seem to believe that models comparable to OPS are all just as bad, RIVM sees, based on the evaluations performed and key figures, that the models (with their own strengths and weaknesses) are all very useful.

But they aren't useful, as we have demonstrated, previously and in this paper. The errors are too large, they mostly don't have skill, and we can't even measure how good they are doing.

However, we heartily agree with RIVM that there are "areas for improvement". More precisely, improvements are necessary as to maintain this model at all. As it stands, OPS should be scrapped until major improvements have been made. The best way to discover improvements are the planned experiments we suggest.

References

[1] W. M. Briggs, J. Hanekamp, and G. Rotgers. Criticizing AERIUS/OPS model performance. https://www.researchgate.net/publication/362578486_Criticizing_AERIUSOPS_Model_Performance.

[2] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *JASA*, 102:359–378, 2007.

[3] T. Gneiting, A. E. Raftery, and F. Balabdaoui. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69:243–268, 2007.

[4] RIVM. Reactie rivm op artikel 'Criticizing AERIUS/OPS model performance' van Briggs, Hanekamp en Rotgers. https://www.rivm.nl/stikstof/actueel/reactie-rivm-op-artikel-criticizing-aeriusops-model-performance. Accessed: 2022-08-23.