Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

Article

Evaluation of Analysis by Cross-Validation. Part I: Using Verification Metrics

Richard Ménard [1]* and Martin Deshaies-Jacques [1]

[1] Air Quality Research Division, Environment and Climate Change Canada; martin.deshaies-jacques@canada.ca
* Correspondence: richard.menard@canada.ca; Tel.: +1-514-421-4613,
2121 Transcanada Highway, Dorval, (QC), CANADA, H9P 1J3

Abstract: We examine how passive and active observations are useful to evaluate an air quality analysis. By leaving out observations from the analysis, we form passive observations, and the observations used in the analysis are called active observations. We evaluated the surface air quality analysis of O_3 and $PM_{2.5}$ against passive and active observations using standard model verification metrics such as bias, fractional bias, fraction of correct within a factor 2, correlation and variance. The results show that verification of analyses against active observations always give an overestimation of the correlation and variance. Evaluation against passive or any independent observations display a minimum variance and maximum correlation as we vary the observation weight, thus providing a mean to obtain the optimal observation weight. For the time and dates considered, the correlation between (independent) observations and the model is 0.55 for O_3 and 0.3 for $PM_{2.5}$ and for the analysis, with optimal observation weight, increases to 0.74 for O_3 and 0.54 for $PM_{2.5}$. We show that bias can be a misleading measure of evaluation and recommend the use of a fractional bias such as the modified normalized mean bias (MNMB). An evaluation of the model bias and variance as a function of model values also show a clear linear dependence with the model values for both O_3 and $PM_{2.5}$.

Keywords: chemical data assimilation; air quality model diagnostics; cross-validation

1. Introduction

Since 2003, Environment and Climate Change Canada (ECCC) has been producing hourly surface analyses of pollutants covering North America [1, 2] which became operational products in February 2013 [3]. The analyses are produced using an optimum interpolation scheme that combines the operational air quality forecast model GEM-MACH output [4] (CHRONOS model output was used prior to 2010 [5]) with real-time hourly observations of O_3, $PM_{2.5}$, PM_{10}, NO_2, and SO_2 from the AirNow gateway with additional observations from Canada. As those surface analyses are not used to initialize an air quality model, it raises the issue on how to evaluate them. We conduct routine evaluations using the same set of observations as those used to produce the analysis. Once in a while, when there is a change in the system, a more thorough evaluation is conducted where we leave out a certain fraction of the observations and use them as independent observations, a process known as cross-validation. Observations used in producing the analysis are called *active observations* while those not used for evaluation are *passive observations*. The purpose of this two-parts paper is to examine the relative merit of using active or passive observations (or independent observations in general) viewed from different evaluation metrics, but also to develop, in Part II, a mathematical framework to estimate the analysis error, and in doing so, to improve the analysis.

The evaluation of an analysis is important, even in the case where it is used to initialize an air quality forecast model, since the evaluation of the resulting air quality forecast may not be a good measure of

the quality of the analysis. In air quality forecasting, the forecast error growth is small, depicts little sensitivity to initial conditions and is in fact more sensitive to numerous modeling errors such as: photochemistry, clouds, meteorology, boundary conditions and emissions just to name a few [6, 7 , 8, 9, 10]. Furthermore, chemical species that are observed are incomplete compared to species needed to initialize an air quality model; incomplete in terms of the number of species observed as well as in their kind [6, 8, 10, 11]. Only a fraction of the observed species (either of secondary or primary pollutants) are usable for data assimilation, important chemical mechanisms are left completely unobserved and for aerosols, information on size distribution is quite limited and almost inexistent when it comes to speciation [6,10]. Also, the observational coverage is limited to the surface or to total column measurements which, up until now, were available at one or two local times per day. There are thus many assumptions to be made from an analysis to a proper 3D initial chemical condition and surface emission correction and its subsequent impact on the air quality forecast. These considerations warrant an independent evaluation of the quality of the analysis on its own [12].

Evaluating an analysis with observations is quite different from evaluating a model with observations, since analyses are created from observations. From a statistical point of view, the observation and analysis cannot be considered independent. However, if we assume that observation errors are spatially uncorrelated. Then, since the passive and active observation sites are never collocated, then the errors from passive observations are uncorrelated with errors of active observations; observations that are used for the analysis. And since the modelling errors is usually assumed to be uncorrelated with observation errors, then it is also uncorrelated with the analysis errors. Cross-validation thus offers a mean to evaluate analyses with statistically independent (passive) observations [13].

In this paper, Part I, we evaluate the relative merit of passive and active observations in the evaluation of analyses using standard metrics used for model evaluation. We show how and when the use of active observations can be misleading and that passive observations can provide a mean to identify optimal analyses. Our examples show that optimal analyses, at the independent observation sites, have much smaller biases than the model biases and increase the correlation coefficient by nearly a factor 2.

The paper is thus organized as follows. First we present the analysis scheme we will be using, as well as the cross-validation design, the evaluation metrics and the configuration of the experiments. Then in §3, we assess the quality of the analyses in both active and passive observation spaces using standard air quality evaluation metrics, identify some pitfalls of some metrics and advocate using active observations. Conclusions are presented in §4.

2. Experimental design

2.1. Design of the objective analysis solver

In optimum interpolation there is no use of an explicit interpolation observation operator. The correlation between a pair of locations, either from two observations sites or from an observation site to a model grid point, is computed as a function of distance using a prescribed correlation function. The observation operator is in effect a delta function applied over a continuous spatial domain [14].

In this study we interpolate the gridded analysis field to observations locations, using bilinear interpolation, to compute residuals such as observation-minus-analysis. Thus there can be a discrepancy between the observation operator used to generate the analysis, i.e. delta functions, and the observation operator used to interpolate the analysis field at the observation location, i.e. bilinear interpolation. To eliminate this discrepancy in observation operators we have revised the optimum interpolation scheme to use explicitly the same bilinear interpolation in handling the error covariance. We will give details below.

As in the operational optimum interpolation, the inversion of the innovation covariance matrix for the analysis solver is done using Choleski decomposition on the full matrix. The number of observations to be processed per analysis being of the order of a thousand or less, there was no need for computational simplification for large number of observations by using either data selection [15] or compact support correlation functions [14, 16]. Thus, the analysis scheme used in this study computes explicitly the gain matrix $\tilde{\mathbf{K}}$ as,

$$\tilde{\mathbf{K}} = \tilde{\mathbf{B}}\mathbf{H}^T(\mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}})^{-1}, \quad (1)$$

where $\mathbf{H}$ is a bilinear interpolation operator, $\tilde{\mathbf{B}}$ is the prescribed background error covariance and $\tilde{\mathbf{R}}$ is the prescribed observation error covariance. The tilde ($\tilde{}$) emphasizes that these are prescribed, potentially suboptimal, quantities.

The computational demand of the Kalman gain was kept low by computing the background error correlation function only at model grid points needed for the bilinear interpolation. For example, to calculate the correlation between a pair of observations requires the computation of correlation between four points surrounding observation 1 (needed for the bilinear interpolation) and the other four points surrounding observation 2, thus forming a 4×4 correlation matrix $\mathbf{C}$ between the target model grid points. Then we calculate $\mathbf{HCH}^T$ which gives the correlation between two observation sites. This procedure is generalized for the N observations needed for the analysis. Equation (1) also involves the computation of $\tilde{\mathbf{B}}\mathbf{H}^T$ that we compute as a set of N representers (i.e. columns of $\tilde{\mathbf{B}}\mathbf{H}^T$), each being a 2D field that maps the background error covariance in model space with a single observation location, using again the bilinear interpolation approach to get a single interpolated representor for each observation location. By doing so we keep the consistency between the observation operators used for interpolation of a field and the observation operator used to manipulate matrices.

2.2. Cross-validation

Cross-validation is a technique to evaluate an analysis (or in general any model that depends on observations) by partitioning the original observation data set into a training set, used to create the analysis, and an independent (or passive) set, used to evaluate the analysis. The most common cross-validation designs are: the *k*-fold cross-validation, where the original observation data set is partitioned into *k* equal size subsamples and the leave-one-out cross-validation, where *N* subsamples are created, each with one different observation set aside for the evaluation while the other $N-1$ observations are used in producing the analysis. The cross-validation is then repeated with all the different sets until all observations have been used for evaluation. Clearly, there are *k* analyses computed in the *k*-fold cross-validation and *N* in the leave-one-out cross-validation, the later being computationally demanding when *N* is large. The main disadvantage of the *k*-fold cross-validation is that the analyses being evaluated uses a smaller number of observations (actually $(k-1)N/k$) than the original observation data set, whereas the leave-one-out cross-validation evaluates analysis that uses nearly the same number of observations (actually $N-1$) as the original observation data set. This actually matters with the *k*-fold cross-validation if we need an estimate of the analysis error variance (or any other second moments) as the analysis error variance depends on the number of observations used.

Let $\mathbf{O}_j$ be a vector that contains the j^{th} set of observations used for evaluation, and let $\mathbf{A}_{(j)}$ be a vector of analysis value interpolated at the verification observation locations of $\mathbf{O}_j$ and where the analysis used all observations except those in $\mathbf{O}_j$ (the index in parenthesis, i.e. (j), indicates all sets except the set *j*). It is customary in cross-validation literature (e.g. [17]) to construct a mean square error cost function, often denoted by CV,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

$$\mathrm{CV} = \sum_j (\mathbf{O}_j - \mathbf{A}_{(j)})^T (\mathbf{O}_j - \mathbf{A}_{(j)}), \quad (2)$$

that represents a misfit quadratic error of the model $\mathbf{A}$ - in our case the analysis. Different model $\mathbf{A}$ can be compared and selected from which the CV value is smallest. Likewise, a tunable parameter in $\mathbf{A}$ can be obtained by minimizing the cost function CV with respect to that parameter. As we shall discuss later in this paper, in §4 and onwards, the bias of $(\mathbf{O}_j - \mathbf{A}_{(j)})$ needs to be removed from the cost function in order to estimate the input error covariance parameters.

In applications and thus in all experiments that follows, the analyses and verification against passive observations are made only with a set of observations that has passed a quality control. The quality control is nearly identical to the quality control used for the operational implementation of the analysis of surface pollutants at ECCC (Robichaud *et al*. [3], supplementary material 1). It consists in discarding observations that report a negative value, or whose value exceeds a certain unrealistic threshold set to 300 ppbv for ozone (300 μg/m^3 for $PM_{2.5}$). Observations are also discarded based on innovations (or observed-minus-background values) when, for ozone, they exceed 50 ppbv (100 μg/m^3 for $PM_{2.5}$) in absolute value. The quality-controlled observations are then separated into 3 sets of observations of equal numbers, i.e. a 3-fold cross-validation procedure, as illustrated in Figure 1.

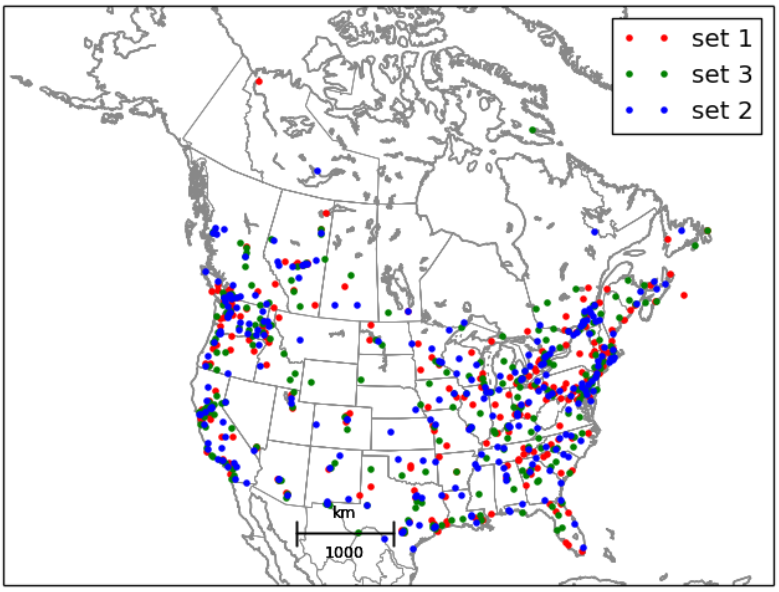

Figure 1. Spatial distribution of the 3 subsets of $PM_{2.5}$ observations used for cross-validation. The selection algorithm is based on regular picking of station by ID number.

The selection into three sets is made by station ID number, selecting on a regular basis each fourth station, starting with station 1 for the first set, station 2 for the second set and station 3 for the third set, and resulting in locally spatially random distribution of each sets of stations. The cross-validation is then made by leaving one set out of the three sets, and using the remaining two sets to produce the analysis.

2.3. Verification metrics

We will evaluate the analyses against passive and active observations with the following standard evaluation metrics used for air quality models [18, 19, 20, 21]; the bias, the modified normalized mean bias (MNMB), the fraction of correct within a factor 2 (FC2), the variance (var(*O*-*A*)) and the correlation

coefficient ($cor(O,A)$), where the statistics is computed over time t for each station, and then the resulting metric is averaged over all the verifying station i,

$$\text{bias} = \frac{1}{N_i}\sum_i\left\{\frac{1}{N_k}\sum_k (O_i(t_k) - A_i(t_k))\right\}, \tag{3}$$

$$\text{MNMB} = \frac{1}{N_i}\sum_i\left\{\frac{2}{N_k}\sum_k\left(\frac{O_i(t_k) - A_i(t_k)}{O_i(t_k) + A_i(t_k)}\right)\right\}, \tag{4}$$

$$\text{FC2} = \frac{1}{N_i}\sum_i\left\{\frac{1}{N_k}\,count\left\{0.5 \le \frac{A_i(t_k)}{O_i(t_k)} \le 2\right\}\right\}, \tag{5}$$

$$\text{var}(O-A) = \frac{1}{N_i}\sum_i\left\{\frac{1}{N_k - 1}\sum_k [(O_i(t_k) - A_i(t_k)) - \overline{(O_i - A_i)}]^2\right\}, \tag{6}$$

$$\text{cor}(O,A) = \frac{1}{N_i}\sum_i\left\{\frac{1}{N_k - 1}\frac{\sum_k (O_i(t_k) - \overline{O}_i)(A_i(t_k) - \overline{A}_i)}{\sqrt{\sum_k (O_i(t_k) - \overline{O}_i)^2 \sum_k (A_i(t_k) - \overline{A}_i)^2}}\right\}. \tag{7}$$

where $O_i(t_k)$ is the observed value at time t_k at the station i, $A_i(t_k)$ is the analysis at time t_k interpolated at the location of the station i, N_k is the total number of time sample per station, N_s is the total number of station (in the sample or over the domain), and the overbar ($\overline{\ }$) denotes the time average. The bias and the MNMB are metrics of the first moment that have distinctive properties. The bias gives a representative measure of the systematic discrepancy between analyzed and observed values over the whole set of observations used for verification. However, since atmospheric constituents exhibit a range of values that can vary in time and space, and that different constituents have different range of values and may as well be expressed with different units, a relative error measure such as the MNMB is often preferred [20]. The MNMB is a dimensionless quantity that falls in the range $[-2,+2]$. The factor 2 is introduced so to give a % error interpretation to the MNMB. This metric has also the additional advantage of treating over- and under-estimation in a symmetric way [21]. However, the MNMB is relatively insensitive to relatively large discrepancies between analysis (or model) values and observed values, that is when its values are close to +2 (200%) or -2 (-200%) [20].

The fraction of correction within a factor 2 (FC2) is a measure of reliability. It is based on counts and has the distinctive advantage that it is insensitive to outliers. It is worth mentioning that it accounts both high values outliers and also low values outliers that is a unique property of this metric [19]. The FC2 metric is also symmetric with respect to permutation of A and O, it is also dimensionless and its values must lie between 0 and 1. Our experience with this metric indicates that it is relatively insentive for relatively good agreement between analysis and observed values.

The variance, var(O-A), and the correlation coefficient $\text{cor}(O,A)$ are metrics that depend on the spread of the discrepancy between analysis and observed values. The variance is not a dimensionless metric. It gives a representative measure of the spread of the discrepancy between analyses and observations and is not sensitive to systematic errors. As we will show in §4 and also shown in Marseille *et al*. [13], var(O-A) with passive observations has the distinct advantage of providing a measure of the true analysis error variance (i.e. the error with respect to the truth) and var(O-A) can be considered as a cross-validation cost function CV, eq.(2), with debiased (O-A) increments. As for any second moment metric, var(O-A) is sensitive to outliers; they must be removed, and this is done by gross check of the $(O-B)$, as explained in the previous subsection §2.2. Finally, the correlation coefficient is a dimensionless quantity that lies in the range [-1, +1]. It is also invariant to shifts in the mean (i.e. not

sensitive to systematic errors), and multiplicative rescaling of either analysis or observations. The correlation is also relatively insensitive to improvement when the correlation is close to 1 or -1.

2.4. Description of the ensemble of analyses and their verification statistics

A series of hourly analyses of O_3 and $PM_{2.5}$ at 21 UTC for a period of 60 days (June 14 to August 12, 2014) were performed with given input error statistics using the operational model GEM-MACH and the real-time AirNow observations as described in the introduction and with quality controlled observations (see subsection §2.2 above). In all experiments, the observation and background error variances, σ_o^2 and σ_b^2, used in the analysis are uniform. The prescribed observation error and background error covariances are given as $\tilde{\mathbf{R}} = \sigma_o^2 \mathbf{I}$, $\tilde{\mathbf{B}} = \sigma_b^2 \mathbf{C}$, where the correlation model $\mathbf{C}$ is a homogeneous isotropic second-order autoregressive model with a correlation length obtained by maximum likelihood, as in Ménard *et al*. [14]. Note that aside from quality control, that ends up rejecting some observations, the analysis uses the observation values and model realizations as is, with no bias correction.

We repeat the series of 60 day analyses for different observation and background error variances chosen in such a way that their sum $\sigma_o^2 + \sigma_b^2$ is equal to $\mathrm{var}(O-B)$ but with different ratios of error variances $\gamma = \sigma_o^2 / \sigma_b^2$. We perform the series of analyses over a wide range of γ ratios in the interval $[10^{-2}, 10^2]$, thus creating on one end analyses with very large observation weights, i.e. $\gamma \ll 1$, such that the analysis interpolated at the active observation sites tend to match the observed value, and on the other end, with $\gamma \gg 1$, creating analyses with very small observation weight producing analyses that are very close to the background (model) state.

The condition $\sigma_o^2 + \sigma_b^2 = \mathrm{var}(O-B)$, called the *innovation variance consistency*, is an important constraint that is useful for the estimation of the *true* error statistics [22]. Indeed, the stronger condition for the full covariance matrices, the *innovation covariance consistency* criterion, takes the form: $< (O-B)(O-B)^T > = \tilde{\mathbf{R}} + \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T$, where $\mathbf{H}$ is the interpolation from model grid to the observation location (or observation operator), and is one of the two necessary and sufficient condition to obtain the *true* error covariance statistics (in observation space) [22, 23].

As explained in the the section §2.3 above, the verification metrics are first calculated over time for each station, i.e. 60 days for a given time, then the metric is averaged over all the verifying stations. If N_s is the total number of stations, the statistics over one of the 3-fold subset then involves an average of the metric over $N_s/3$ passive stations. Doing this for all 3 subsets, and taking the average of the subsets' results, is equivalent to taking the average of the metric over all stations. In the results that will be presented in the following sections, we always present the average metric over the 3 passive subsets so that, in the end, the sample size of the passive observation experiments and of the active observation experiments are equal and thus can be presented side by side on the same graphic.

3. Verification against passive and active observations

In this series of experiments, analyses of O_3 and $PM_{2.5}$ were produced using a fixed homogeneous isotropic correlation function, where the correlation length was obtained by maximum likelihood using a second-order auto-regressive model and error variances computed using a local Hollingsworth-Lönnberg fit [14]. A correlation length of 124 km was obtained for O_3 and of 196 km for $PM_{2.5}$. Our correlation length is defined from the curvature at the origin as in Daley [24] and is different from the length-scale parameter of the correlation model (see Ménard *et al*. [14] for a discussion of these issues). We did a series of 60-days analyses for different values of σ_o^2 and σ_b^2 but such that their sum respects the innovation variance consistency, $\sigma_o^2 + \sigma_b^2 = \mathrm{var}(O-B)$, an important condition for an optimal analysis [22], as explained in §2.4. The results are shown for a wide range of variance ratios $\gamma = \sigma_o^2 / \sigma_b^2$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

from 10^{-2} to 10^{2}. Note that $\gamma \ll 1$ corresponds to a very large observation weight while $\gamma \gg 1$ correspond to very small observation weight.

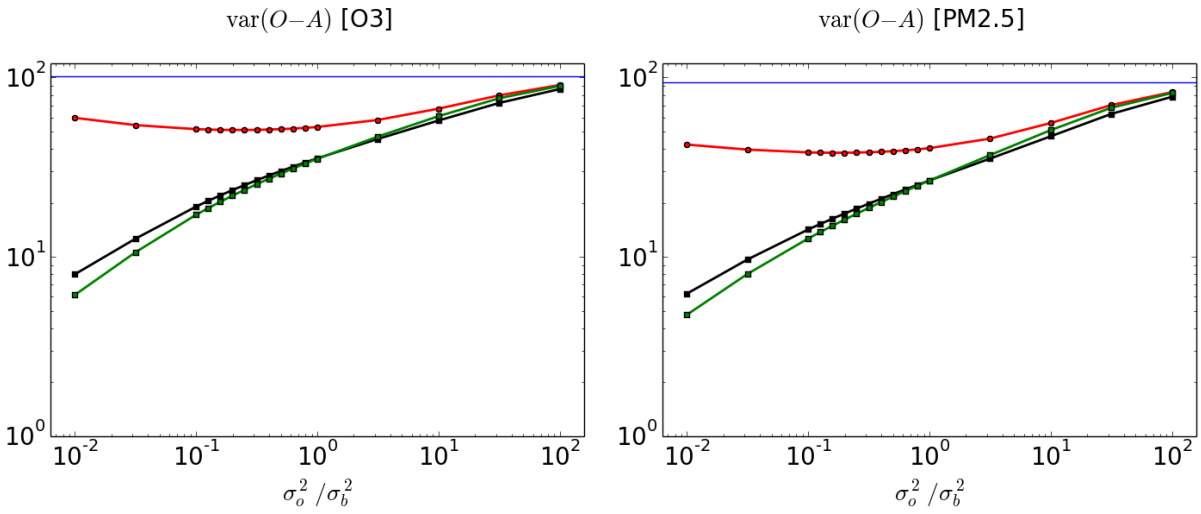

Figure 2. Variance of observation-minus-analysis residuals of O_3 and $PM_{2.5}$ for both active and cross-validation passive observations as a function of $\gamma = \sigma_o^2/\sigma_b^2$. Left panel is for O_3 with ordinates in ppbv^2 units, and right panel is for $PM_{2.5}$ with ordinates in $(\mu g/m^3)^2$. Red curve results from the evaluation at the passive observation sites (average of the 3-fold subsets). Black curves results from evaluation at the active observation sites with analyses using all observations. Green curves results from the evaluation at the active observation sites in the cross-validation experiment (i.e. using 2/3 of the observations; average of the 3 subsets). Blue curve is the variance of observation-minus-model.

The $\mathrm{var}(O-A)$ using passive observations (red curve with circles) and active observations (black curve with squares) is presented in Figure 2 for O_3 (left panel) and $PM_{2.5}$ (right panel). The solid blue line represents $\mathrm{var}(O-B)$, the variance of observation-minus-model, i.e. prior to an analysis. As mentioned in §2.4, in the cross-validation experiments we averaged the verification metric over the 3-fold subsets so that, in effect, the total number of observations that ends up being used for verification is N_s, the total number of stations. We thus argue that the verification sampling error for the cross-validation experiments (red curve) is the same as for the active observations using the full analysis (i.e. analysis using the total number of stations; black curve). Also there is roughly 1,300 quality controlled O_3 observations over the domain and 750 $PM_{2.5}$ quality controlled observations, each with 60 time samples or less. To give some qualitative idea of the sampling error, the different metric values for the individual 3-fold sets are presented in the supplementary material section, where we can see that for $\mathrm{var}(O-A)$ and $\mathrm{cor}(O,A)$ the metric values for the individual sets are nearly indistinguishable from the means of the 3-subset.

The difference between the verification against passive observations in cross-validation analyses (red curve) and the verification against active observations using full analyses (black curve) can be attributed to two effects: 1- the analysis used in the cross-validation uses 2/3rd of the total number of observations and thus the analysis error has larger variance than analyses using all observations, 2- there is a distance effect between the passive observation sites and the active ones for the analyses using 2/3rd of the observations. In order to separate these two effects, we also display the 3-fold average of the metric verifying against active observation for the cross-validation analyses as a green curve with squares. Thus in summary we display a;

- red curve : using analysis with $2N_s/3$ observations with an evaluation at passive sites
- green curve: using analysis with $2N_s/3$ observations with an evaluation at active sites
- black curve : using analysis with N_s observations with an evaluation at active sites.

The difference between the red and green curves show the influence of distance between passive and active observation sites, whereas the difference between the green and black curves show the influence of having different number of observations in creating the analysis for verification.

Now let us examine the results of verifying against passive observations with cross-validation analyses. As the observation weights gets smaller (i.e. $\gamma \gg 1$), the analysis draws closer to the background, so that $\mathrm{var}(O-A)$ increases toward $\mathrm{var}(O-B)$. On the other end when $\gamma \ll 1$, the analysis tries to overfit active observations which results in a spatially noisy analysis, which explains that $\mathrm{var}(O-A)$ increases as γ diminishes. Somewhere in between lies a minimum of $\mathrm{var}(O-A)$ where there is neither an overfitting nor an underfitting to the active observations. This "optimal" ratio actually corresponds the optimal analysis as we shall discuss in Part II of this study.

Now examining the results of verifying against active observations gives a different message. The verification against active observations is presented with the black curves for the full analyses and with the green curves for the cross-validation sets; the difference between the two curves being the number of observations used to generate the analyses. In both curves we observe that $\mathrm{var}(O-A)$ is steadily decreasing as the observation weigth increases. In effect, it is an expected result from the inner working of an analysis scheme that the analysis error variance does not depend on the observed values or the model values. For this reason, the $\mathrm{var}(O-A)$ using active observations cannot provide a true measure of the quality of an analysis. There is, however, an exception to this when the analysis is optimal as we shall see in Part II of this paper.

One would expect from having a larger number of observations in the analysis that the $\mathrm{var}(O-A)$ for the cross-validation analyses be slightly smaller than the $\mathrm{var}(O-A)$ for the full analysis. This is observed between the black and green curves when the observation weight is small (i.e. $\gamma \gg 1$). However, and surprisingly when the observation weight is large, $\gamma \ll 1$, we observe the opposite. This intriguing behavior may indicate an inconsistency between the assumption of uniform error variances for σ_o^2 and σ_b^2 (assumed in the input error statistics) and the real spatial distribution of error variances. This discrepancy being simply amplified when the observation weight is large and when there are less observations to produce the analysis.

The difference between $\mathrm{var}(O-A)$ at passive sites and active sites (with the same number of observations to construct the analyses) is significant. For O_3 and for an optimal ratio, the $\mathrm{var}(O-A)$ at passive sites is 51.02 ppbv^2 (red curve) while at active sites is 22.77 ppbv^2 (green curve). For PM2.5 and for an optimal ratio, the $\mathrm{var}(O-A)$ at passive sites is 38.09 $(\mu g/m^3)^2$ (red curve) while at active sites is 15.41 $(\mu g/m^3)^2$ (green curve). For both species, the error variance at active sites gives a significant overestimation of the error variance by more than a factor 2.

In Figure 3, we present the correlation metric between the observations and the analysis using, as in Figure 2, the verification against passive observations in cross-validation analyses (red curve), the verification against active observations using full analyses (black curve) and the verification against active observations in the cross-validation analyses (green curve). The blue curve depict the correlation between the model and the observations, that is the prior correlation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

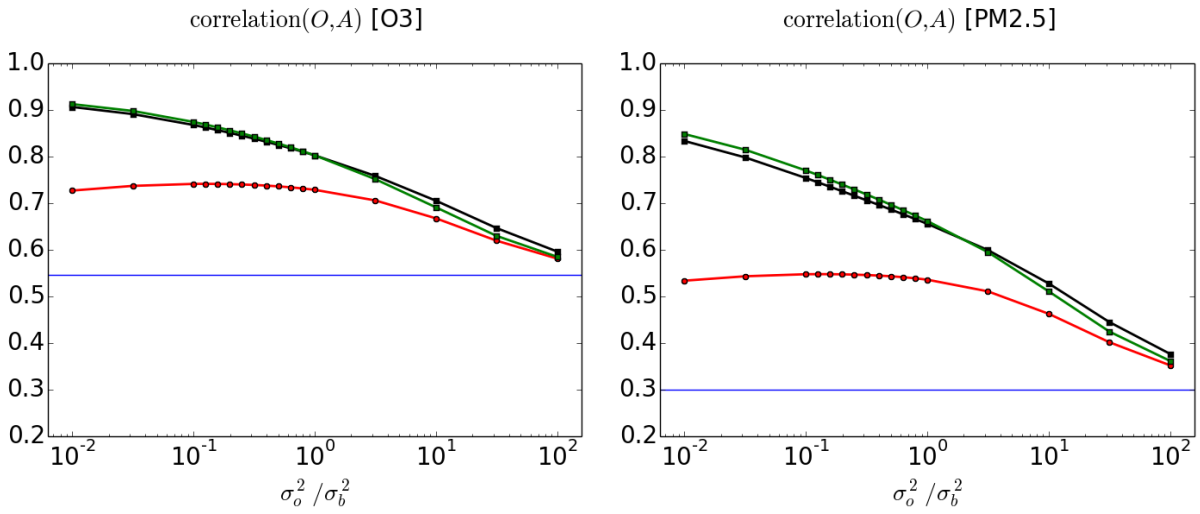

Figure 3. Correlation between observations and analysis for O_3 and $PM_{2.5}$ for both active and cross-validation passive observations as a function of $\gamma = \sigma_o^2 / \sigma_b^2$. The red, black and green curves are as in Figure 2.

The evaluation against passive observations with cross-validation analyses (red curve) shows a maximum at the same values of $\gamma = \sigma_o^2 / \sigma_b^2$ than for the $\mathrm{var}(O - A)$. We argue that the same arguments of underfitting and overfitting are responsible for this maximum. The correlation between the active observations and the analysis (black and green curves) increases as the observation weight increases (γ decreases), theoretically reaching a value 1 for $\sigma_o^2 = 0$, which is again unrealistic and simply shows the impact of ill-prescribed error statistics in an analysis scheme. The gain in correlation between independent observations and analysis is significant. For O_3, it increases from a value of 0.55 with respect to the model to a value of 0.74 with respect to an optimal analysis (when $\gamma = \sigma_o^2 / \sigma_b^2$ is optimal). For $PM_{2.5}$, the correlation against the model has a value of 0.3 which basically has no skill, to a value of 0.54 for optimal analysis, which represent a modest but useable skill. The correlation evaluated at the active sites for an optimal ratio, is 0.85 for O_3 (green curve) and 0.74 for $PM_{2.5}$ (green curve), again being a significant overestimation with respect to values obtained at passive sites.

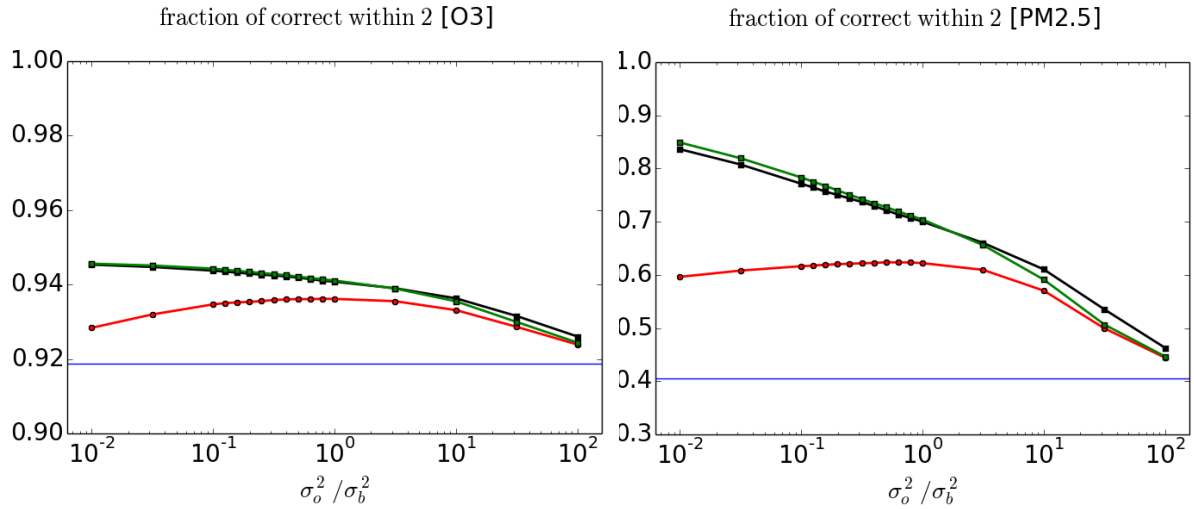

Figure 4. Fraction of correct within a factor 2 for O_3 (left panel) and $PM_{2.5}$ (right panel) for both active and cross-validation passive observations as a function of $\gamma = \sigma_o^2 / \sigma_b^2$. The red, black and green curves are as in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

Another metric that we have considered is the fraction of correct within a factor 2, eq. (5) [3]. The evaluation of this metric against passive and active observations is presented in Figure 4 for O_3 (left panel) and $PM_{2.5}$ (right panel). Note that the scale in the ordinate is quite different between the left and right panels. Although the results bear similarity with the correlation between O and A presented in Figure 3, the maximum with passive observations is reached at larger γ values than those obtained for $\mathrm{var}(O-A)$ or $\mathrm{cor}(O,A)$, which are identical. One possible explanation is that biases in observations and analysis are not removed in the metric which could explain the shift of the maximum.

The interpretation of this metric is, however, not clear. Although the ratio $z=A/O$ is a dimensionless quantity the spread of z is generally not independent of the variance of A or O and there are cases where it is. So to count the number of occurrence of z between the dimensionless values 0.5 and 2 is confusing. As a simplified illustration, suppose that A is normally distributed as $N(0,\sigma_a^2)$ and similarly with $O \sim N(0,\sigma_o^2)$. The ratio of these two random variables is then a Cauchy distribution whose probability density function (pdf) is $\sigma_o\sigma_a/[\pi(\sigma_o^2 z^2+\sigma_a^2)]$. The mean, variance and higher moments of Cauchy probability distributions are not defined since the integral of the pdf is not bounded; only the mode is defined. Cauchy distributions also have a spread parameter, which in this case is equal to σ_a/σ_o. If the variance of A and O are equal, then the number count between the dimensionless bounds 0.5 and 2 depends only on the shape of the probability distribution function, not on the variance. If the variance of A and O are different, then it also depends on the ratio of variances. Furthermore, this metric also depends on the bias. It is a difficult metric to interpret and overall it is unclear what new information it adds. If used as a quality control, however, the FC2 have the unique ability of rejecting too low as well as too high values of z.

In Figure 5 we present the bias between observations and analyses, and were the verification is made against passive and active observations as done with the other metrics. Bias is not a dimensionless quantity; note that the range and scale presented for O_3 and $PM_{2.5}$ in Figure 5 are different. The blue curve is the mean $(O-B)$ and thus indicates for O_3 in average over all observation stations (for the time and dates considered) the model overpredicts, and that for $PM_{2.5}$ the model underpredicts.

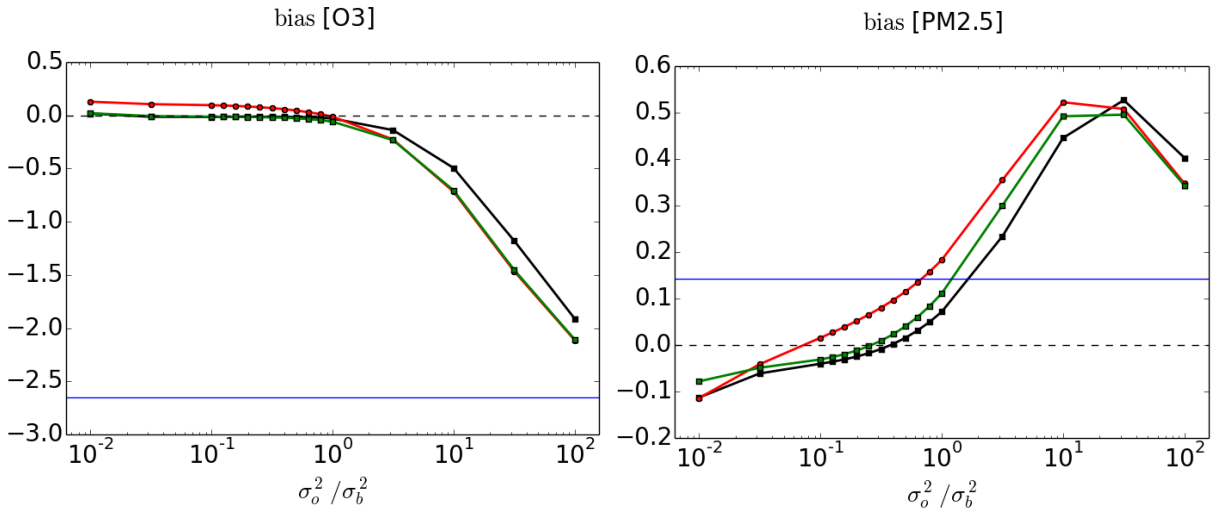

Figure 5. Bias between observation and analysis for O_3 (left panel) and $PM_{2.5}$ (right panel) for both active and cross-validation passive observations as a function of $\gamma=\sigma_o^2/\sigma_b^2$. The red, black and green curves are as in Figure 2.

Contrary to all metric results seen so far, the sampling uncertainty of the bias is significant: it is of the order of ± 0.5 ppbv (in average) for O_3 at passive sites and of the order of $\pm 0.1\ \mu g/m^3$ (in average) for the $PM_{2.5}$ at passive sites (results shown in the supplementary material). The distinction between the red, black and green curves is thus not significant for both O_3 and $PM_{2.5}$. But, the difference between the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

analysis bias and model bias is nearly always significant. For O_3, the model bias is eliminated at the passive observation sites (red curve) as long as the observation weight $\gamma \leq 1$. The situation is not so clear for $PM_{2.5}$. In fact, when the observation weight is small, the bias result indicates that the analysis has a larger bias than the model. How can that be when the observation weight is small (i.e. $\gamma > 1$) and thus the analysis should be close to the model values? This apparent contradiction reveals a more complex issue underlying the bias metric.

To explore the possible causes, we have calculated the bias per bin of model values, displayed in Figure 6.

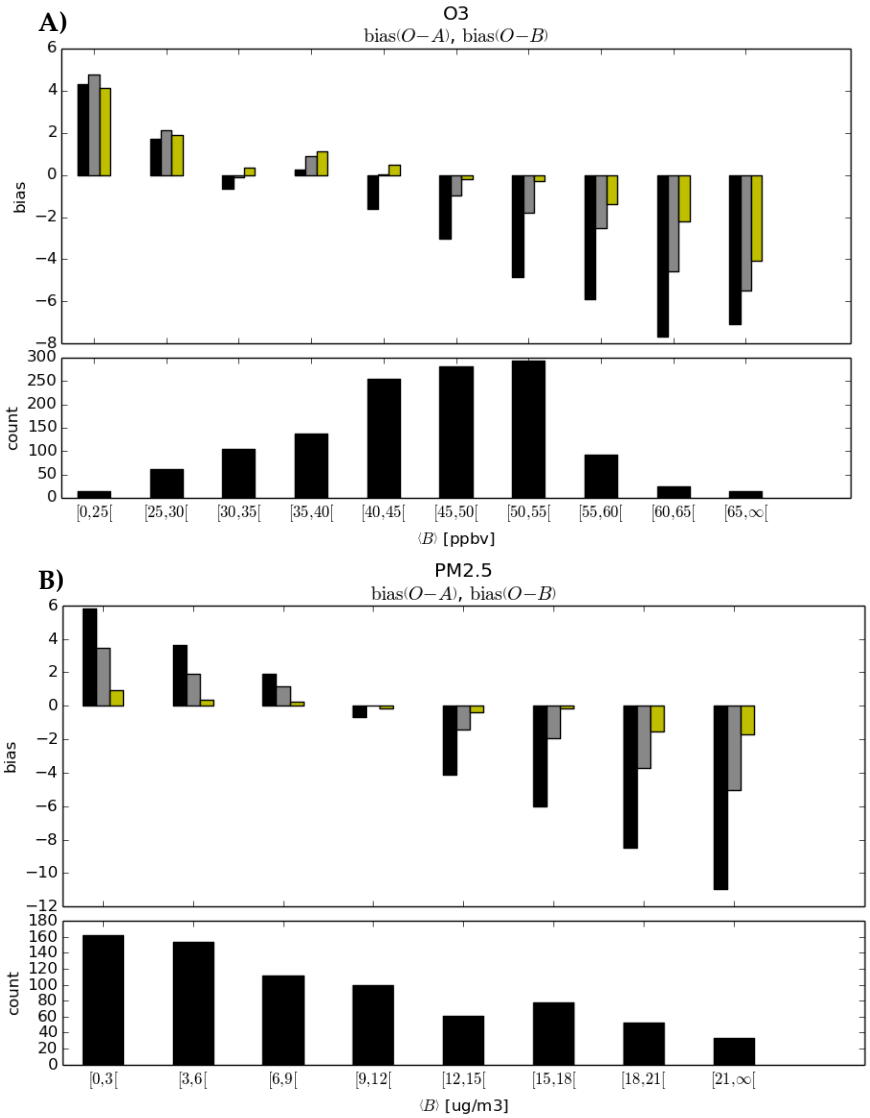

Figure 6. Biases per bin of model values. In **A)**, presents the statistics for O_3 and in **B**), for $PM_{2.5}$. In the upper portion of figures **A)** and **B)** are the residual statistics per bin; in black, the $(O-B)$, in grey, the $(O-A)$ at passive observation sites (mean of the 3-fold subsets) for a non-optimal analysis with $\gamma = 10$, and in yellow, the $(O-A)$ at passive observation sites (mean of the 3-fold subsets) using the optimal observation weight. The lower portion of the figures **A)** and **B)** are the station number count per model values.

In order have decent sample size per bin, we collect all the $(O-A)$ and $(O-B)$ over time and observation sites, create bins of model values and calculate the statistic per bin (and not per stations as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

before). The result shows that the model bias is nearly linearly dependent on the model (black boxes in the bias panel). Both O_3 and $PM_{2.5}$ show an underprediction for low model values and an overprediction for large model values. The origin of this bias is not known but one would argue that it is not related to chemistry since both constituents, O_3 and $PM_{2.5}$, present the same feature. In the lower panels of **A)** and **B)** shows the count of stations per model bin size. We observe that the majority of stations have O_3 model values in the range of 40 to 55 ppbv, where the bias is negative. Over all the stations, this give rise to a negative mean $(O-B)$, and this is how we claim that claim that the model overpredicts. However, for $PM_{2.5}$ the situation is different: the majority of stations lie in the low model value range, and there are gradually less stations for increasingly larger model values. Although the $(O-B)$ have large negative values in the high model value bin while small model value bins have positives $(O-B)$'s, the effect over all stations is to yield a modestly positive mean $(O-B)$ and thus the model underestimates the $PM_{2.5}$. The results of the analysis evaluated at the passive observation sites are presented with the yellow and grey histogram boxes. In yellow, near optimal analyses with optimal observation weight, as determined by the minimum of $\mathrm{var}(O-A)$ are used, and in grey non-optimal analyses with $\gamma = 10$. We observe that the effect of the optimal analysis is nearly insentive to model bin values, where near zero biases are obtained in most of the range except for very small and very large model values. The fact that we are not able to capture the full benefit of analysis on all model values may be an artefact of the assumption that we are using uniform observation and background error variances whereas the model values varies considerably. In grey, we used the non-optimal analyses with a small observation weight were we set $\gamma = 10$. In the non-optimal case, the state-dependent bias is still present but appears to be nearly perfectly anti-symmetric, positive in the low model value bins and nearly exact opposite in high model value bins. Since for O_3 the majority of observations lies in the range 40 to 55 ppbv, $(O-A)$ for the optimal analyses at passive observation sites is nearly zero. But, for the non-optimal analysis with $\gamma = 10$, the $(O-A)$ at passive sites is negative, i.e. the analysis is overpredicting, as shown in Figure 5.

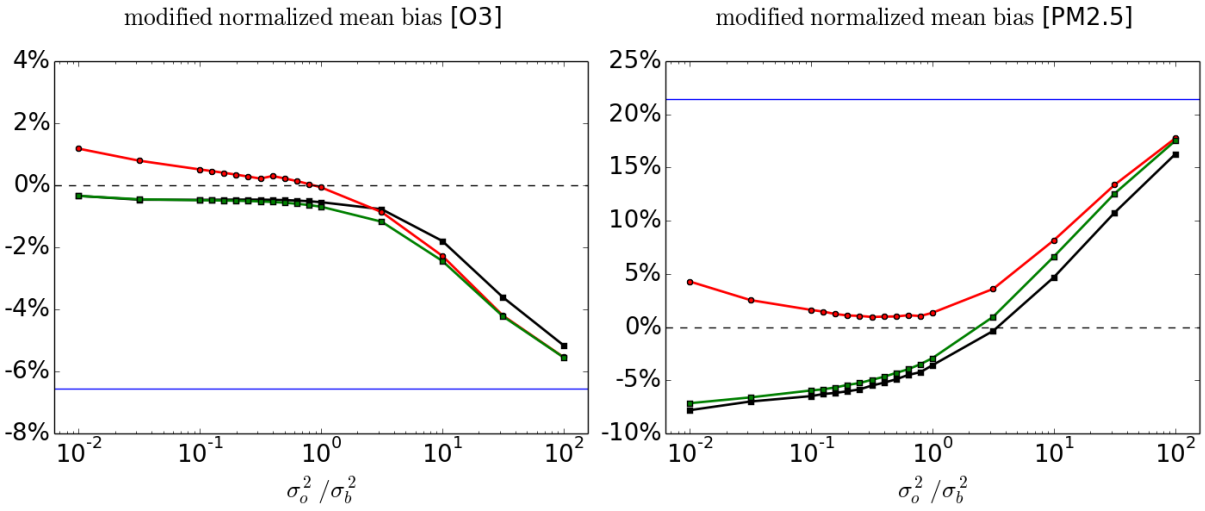

Figure 7. Modified normalized mean bias (MNMB) between observation and analysis for O_3 (left panel) and $PM_{2.5}$ (right panel) for both active and cross-validation passive observations as a function of $\gamma = \sigma_o^2/\sigma_b^2$. The red, black and green curves are as in Figure 2.

For $PM_{2.5}$, the weighted sum of the $(O-A)$ bins is such that over all stations the bias for an optimal analysis is nearly zero. In the case of the non-optimal analysis with $\gamma = 10$, the weighted sum of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 January 2018 doi:10.20944/preprints201801.0217.v1

Peer-reviewed version available at Atmosphere 2018, 9, 86; doi:10.3390/atmos9030086

nearly anti-symmetric $(O-A)$ bias per bin gives more weight to the positive bias at smaller model values, so that overall there is a positive $(O-A)$, as in Figure 5.

To circumvent the state-dependency of the $(O-A)$ biases it is useful to consider instead a fractional bias metric, such as the modified normalized mean bias, MNMB eq.(4). The MNMB metric is a dimensionless measure and as defined with a factor 2, eq.(4), represents a % error. The MNMB for O_3 and $PM_{2.5}$ for passive and active observations are displayed in Figure 7 using the same color as in Figure 2. We note immediately that the MNMB analysis bias does not exceed the MNMB model bias as we observed for the bias metric of $PM_{2.5}$ (Figure 5 right panel). The MNMB bias also varies smoothly as a function of γ (at variance with the bias metric for $PM_{2.5}$ – Figure 5).

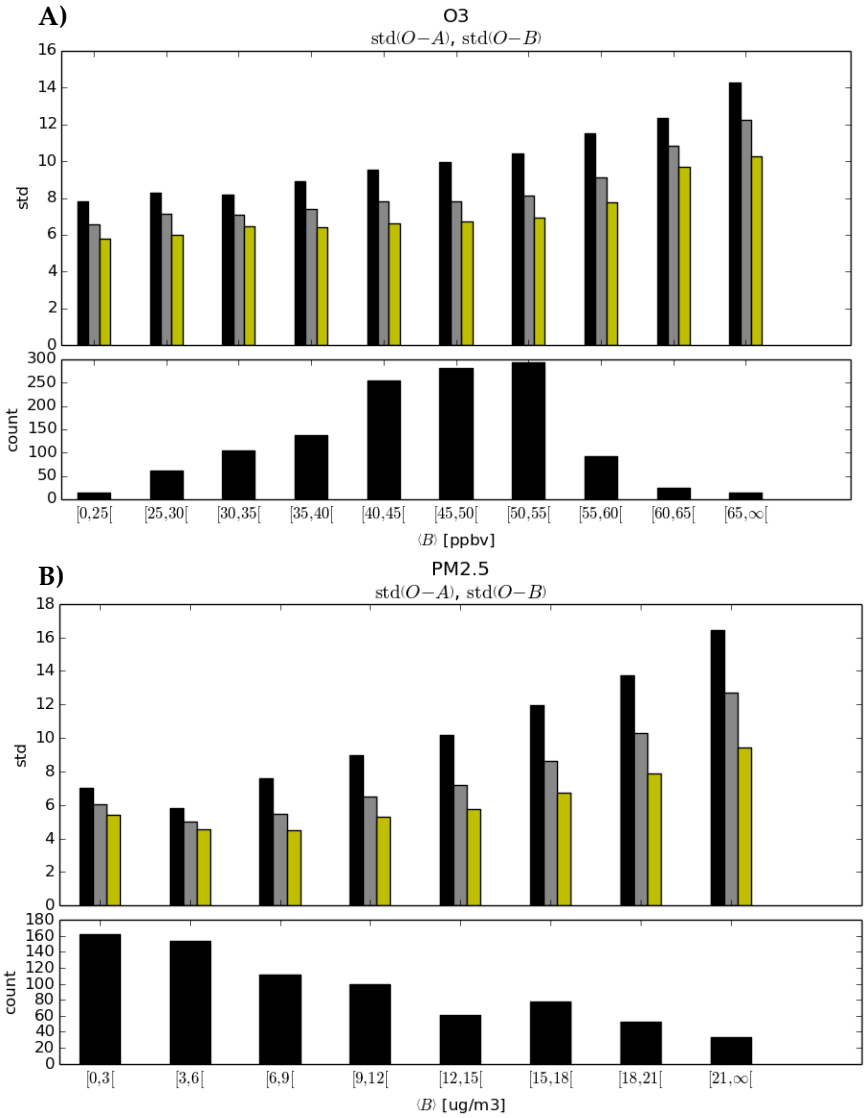

Figure 8. Same as Figure 6 except that we display the variance of analysis-minus-passive observations per bin of model values.

Furthermore the sampling uncertainty of the analysis MNMB relative to the model bias (see the supplementary material) is smaller than the relative uncertainty of the analysis bias with respect to the model bias. This is especially true for $PM_{2.5}$, where we can actually deduce that the difference between the cross-validation and the validation against active observations is significant when $\gamma < 1$. There is also another important point to make; although analyses are designed to reduce the error variance, it

thus so happens that for a near optimal analysis the fractional bias MNMB is very small, around 1% for O_3 and about 1-2% for $PM_{2.5}$. We argue that it results from an optimal use of observations.

There is also some information to gain from the variance of observed-minus-analysis per bin size, as illustrated in Figure 8, using the same color histograms as in Figure 6. We note that for O_3, the model error variance against observations increases gradually with larger model values. But the fraction of analysis variance vs. model variance is roughly uniform across all bins. This can be explained by the fact that the observation and background error variances are uniform, and thus the reduction of variance across all bins is uniform as well. However, the situation is different for $PM_{2.5}$. We note a relatively poor performance of the model at low model values, with standard deviation of 7μg/m^3. For slightly larger model values (3-6 μg/m^3), the error variance is smaller to 5.5 μg/m^3 and then increases almost linearly with model values. The fraction of analysis variance vs. model variance decreases steadily with larger model values. These results thus indicate that the assumption that observation and background error variances are uniform and independent of the model value may have to be revisited.

4. Conclusions

We have developed an approach by which analyses can be evaluated and optimized without using a model forecast but rather by partitioning the original observation data set into a training set, to create the analysis, and an independent (or passive) set, used to evaluate the analysis. This kind of evaluation by partitioning is called cross-validation.

The need for such a technique came about our desire to evaluate our operational surface air quality analyses that are created off-line with no assimilation cycling. Evaluating a surface air quality analysis based on its chemical forecast does in fact require additional information or assumptions, such as vertical correlation, aerosol speciation and bin distribution (while surface measurement is primarily about mass) or unobserved chemical variables correlations, and so on…. So that the quality of the chemical forecast is not solely dependent on the quality of the analysis and, if there are compensating errors, can actually be a misleading assessment of the quality of the analysis.

We have applied this cross-validation procedure to the operational analyses of surface O_3 and $PM_{2.5}$ over North America for a period of 60 days and present an evaluation using different metrics; bias, modified normalized mean bias, variance of observation-minus-analysis residuals, correlation between observation and analysis, and fraction of correction within a factor 2.

Our results show that, in terms of variance and correlation, the verification of analyses against active observations always yield an overestimation of the accuracy of the analysis. This overestimation also increases as the observation weight increases. On the other hand for biases, the distinction between the verification against active observations and passive observations is unclear and drowned in the sample variability. However, using a fractional bias metric, in particular the MNMB, shows that the verification against passive observations can be close to one percent for an optimal analysis while the verification against active observations is much larger.

Results also show the importance of having an optimal analysis for verification. The variance of the analysis with respect to independent observations is minimum and the correlation between the analysis and independent observations is maximum for an optimal analysis. By being a compromise between an overfit to the active observations (which produce noisy analysis field) and an underfit, the optimal analysis offers the best use of observations throughout. At optimality, the analysis fractional bias (MNMB) at the passive observation sites has only one or two percent error whereas the fractional bias of the model is 6.5% for O_3 and 21% for $PM_{2.5}$. The correlation between the analysis and independent observations is also significantly improved with an optimal analysis: the correlation between the model and independent observations is 0.55 for O_3 and increases to 0.74 with the analysis, while for $PM_{2.5}$ the correlation between the model and independent observations is only 0.3 (which is basically no skill) but rises to 0.54 for the analysis.

We also argue that the fraction of correct within a factor 2, is a metric whose interpretation is unclear as it mixes information about bias, variance and probability distribution in a non-uniform way and does not seem to add anything new to other metrics. The bias is also very sensitive to sample variability and can lead to wrong conclusions. For example, we have seen that the mean analysis bias can be larger than the mean model bias, whether verifying against active or passive observations. But, since an analysis is always closer to the truth than its prior (i.e. the model), it results in an apparent contradiction. This implies that the bias metric cannot be used to faithfully compare model states accurately. Such wrongful conclusions do not arise, however, with the MNMB. We thus recommend avoiding using bias as a measure of truthfulness, and use instead a fractional bias measure such as the MNMB.

We also found that errors in the GEM-MACH model grow almost linearly with the model value. This is particularly evident for the bias where the model underestimates at small model values and overestimates at large model values. Furthermore this occurs in equal ways for O_3 and $PM_{2.5}$, thus indicating that the source of this bias is not related to chemistry. The fact that, over the entire domain, the model overestimates O_3, and underestimates $PM_{2.5}$ is simply a result of the concentrations. We have not conducted a systematic study of model error for other times of the day and other periods of the year, but it would be very interesting to look at this, to see whether or not changes of biases are due primarily to changes in the distribution of values rather than a fundamental change in the bias per model value bin.

Finally we have also examined the variance against independent observations per model value bin , and concluded that the error variance is not quite uniform with model values but increases slowly with model values for O_3 and in a more pronounced way for $PM_{2.5}$.

In Part II, we will focus on the estimation of the analysis error variance and develop a mathematical formalism that permits to compare different diagnostics of variance under different assumptions, optimize the analysis parameters and gain confidence on the estimate of analysis error as we obtain coherent estimated values across different diagnostics.

Supplementary Materials: The following are available online at www.mdpi.com/link, Figure S1: Verification of variance for O_3 and $PM_{2.5}$ for the individual sets. Figure S2: Same as Figure S1 but for the correlation between observations and analysis. Figure S3: Same as Figure S1 but for the fraction of correct within a factor 2. . Figure S4: Same as Figure S1 but for bias. . Figure S5: Same as Figure S1 but for modified normalized mean bias.

Acknowledgments: We are grateful to the US/EPA for the use of the AIRNow database for surface pollutants and to all provincial governments and territories of Canada for kindly transmitting their data to the Canadian Meteorological Centre to produce the surface analysis of atmospheric pollutants. We are also thankful for the proof read by Kerill Semeniuk, and for two anonymous reviewers for their comments and help in improving the manuscript.

Author Contributions: This research was conducted as a joint effort by both authors. R.M. contributed to the theoretical development and wrote the paper, and M.D.-J. conducted all experiments design and execution, proof reading and introduced a new diagnostic of optimal analysis error that was furher extended to passive observations space.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors, which is the government of Canada, had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Ménard, R., and A. Robichaud. The chemistry-forecast system at the Meteorological Service of Canada. In *ECMWF Seminar Proceedings on Global Earth-System Monitoring*, September 5–9, **2005**, Reading, UK, 297–308.

2. Robichaud, A. and R. Ménard. Multi-year objective analysis of warm season ground-level ozone and PM2.5 over North-America using real-time observations and Canadian operational air quality models. Atmos. Chem. Phys. **2014**. 14:1769-1800. DOI: 10.5194/acp-14-1769-2014.
3. Robichaud, A., R. Ménard, Y. Zaïtseva, and D. Anselmo. Multi-pollutant surface objective analyses and mapping of air quality health index over North America. Air Qual. Atmos. Health. **2016**, DOI 10.1007/s11869-015-0385-9.
4. Moran, M.D., S. Ménard, R. Pavlovic, D. Anselmo, S. Antonopoulus, A. Robichaud, S. Gravel, P.A. Makar, W. Gong, C. Stroud, J. Zhang, Q. Zheng, H. Landry, P.A. Beaulieu, S. Gilbert, J. Chen, and A. Kallaur. Recent advances in Canada's national operational air quality forecasting system. 32nd NATO-SPS ITM, Utrecht, NL, 7-11 May **2012**.
5. Pudykiewicz, J.A., A. Kallaur, and P.K. Smolarkiewcz. Semi-lagrangian modelling of tropospheric ozone. Tellus, 49B, **1997**, 231-248.
6. Carmichael, G.R., A. Sandu, T. Chai, D.N. Daescu, E.M. Constantinescu and Y. Tang. Predicting air quality : Improvements through advanced methods to integrate models and measurements. *J. Comput. Phys*., 227, **2008**, 3540-3571.
7. W.F. Dabberdt, M.A. Carroll, D. Baumgardner, G. Carmichael, R. Cohen, T. Dye, J. Ellis, G. Grell, S. Grimmond, S. Hanna, J. Irwin, B. Lamb, S. Madronich, J. McQueen, J. Meagher, T. Odman, J. Pleim, H.P. Schmid, D.L. Westphal, Meteorological research needs for improved air quality forecasting – Report of the 11th prospectus development team of the US weather research program, Bull. Amer. Meteorol. Soc. 85 (4), **2004**, 563.
8. Sportisse, B.. A review of current issues in air pollution modeling and simulation. *Comput. Geosci*., **2007**, 11:159-181. DOI 10.1007/s10596-006-9036-4.
9. Elbern, H., A. Strunk, and L. Nieradzik. Inverse modelling and combined state-source estimation for chemical weather. In Data Assimilation: Making Sense of Observation (eds. Lahoz, W, B. Khattatov, and R. Ménard), Springer, **2010**, 491-513.
10. Bocquet, M., H. Elbern, H. Eskes, M. Hirtl, R. Žabkar, G.R. Carmicheal, J. Flemming, A. Inness, M. Pagaoski, J.L. Pérez Camaño, P.E. Saide, R. San Jose, M. Sofiev, J. Vira, A. Baklanov, C. Carnevale, G. Grell, and C. Seigneur. Data assimilation in atmospheric chemistry models; current status and future prospects for coupled chemistry meteorology models. Atmos. Chem. Phys., 15, 5325-5358, **2015**, www.atmos-chem-phys-net/15/5325/2015/, doi: 10.5194/acp-15-5325-2015.
11. T. Chai, G.R. Carmichael, A. Sandu, Y.H. Tang, D.N. Daescu, Chemical data assimilation of transport and chemical evolution over the pacific (TRACE-P) aircraft measurements, J. Geophys. Res. 111 (D02301), **2006**, doi:10.1029/2005JD005883.
12. Sandu, A., and T. Chai. Chemical data assimilation–An overview. *Atmosphere* **2011**, 2, 426-463; doi:10.3390/atmos2030426.
13. Marseille, G.-J., J. Barkmeijer, S. de Haan, and W. Verkley. Assessment and tuning of data assimilation systems using passive observations. *Q. J. R. Meteorol. Soc.*, **2016**, 142, 3001-3014, DOI:10.1002/qj.2882.
14. Ménard, R., M. Deshaies-Jacques, and N. Gasset. A comparison of correlation-length estimation methods for the objective analysis of surface pollutants at Environment and Climate Change Canada. *Journal of the Air & Waste Management Association*. **2016**, 66:9, 874-895, DOI: 10.1080/10962247.2016.1177620.
15. Cohn, S.E., A. da Silva, J. Guo, M. Sienkiewicz, and D. Lamich. Assessing the effects of data selection with the DAO physical-space statistical analysis system. *Mon. Wea. Rev*. **1998**, 126, 2913–2926.
16. Houtekamer, P.L., and H.L. Mitchell. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev*. **2001**, 129,123-137.
17. Efron, B. An introduction to boostrap. Chapman & Hall: New York, **1993**, 436 pp.
18. Seigneur, C., B. Pun, P. Pai, J.-F. Louis, P. Solomon, C. Emery, R. Morris, M. Zahniser, D. Worsnop. P. Koutrakis, W. White and I. Tombach. Guidance for the performance evaluation of three-dimensional air quality modeling systems for particulate matter and visibility. Journal of the Air & Waste Management Association. **2000**, 50:4, 588-599, DOI: 10.1080/10473289.2000.10464036 .
19. Chang, J.C. and S.R. Hanna. Air quality model performance evaluation. Meteorol. Atmos. Phys.. **2004**, 87, 167-196, DOI 10.1007/s00703-003-0070-7.

692 20. Savage, N.H., P. Agnew, L.S. Davis, C. Ordóñez, R. Thorpe, C.E. Johnson, F.M. O'Connor, and M. Dalvi. Air
693 quality modelling using the Met Office Unified Model (AQUM OS24-26): model description and initial
694 evaluation. Geosci. Model Dev.. **2013**, 6, 353-372, www.geosci-model-dev.net/6/353/2013/
695 21. Katragkou, E., P. Zanis, A. Tsikerdekis, J. Kapsomenakis, D. Melas, H. Eskes, J. Flemming, V. Huijnen, A. Inness,
696 M.G. Schultz, O. Stein, and C.S. Zerefos. Evaluation of near surface ozone over Europe from the MACC
697 reanalysis. Geosci. Model Dev. **2015**, 8, 2299-2314, https://doi.org/10.5194/gmd-8-2299-2015
698 22. Ménard, R. Error covariance estimation methods based on analysis residuals: theoretical foundation and
699 convergence properties derived from simplified observation networks. Q. J. R. Meteorol. Soc.. **2016**, 142, 257-
700 273, DOI:10.1002/qj.2650.
701 23. Desroziers, G., L. Berre, B. Chapnik, and P. Poli. Diagnosis of observation-, background-, and analysis-error
702 statistics in observation space. Q. J. Roy. Meteorol. Soc. **2005**, 131, 3385-3396.
703 24. Daley, R. *Atmospheric Data Analysis*. Cambridge University Press: New York, **1991**, 457 pp.