Optimal Estimation

Understanding Sources of Error When Retrieving Atmospheric Variables

Adapted from "Optimal Estimation Retrievals and Their Uncertainties: What Every Atmospheric Scientist Should Know," by **Maximilian Maahn** (University of Colorado and NOAA/Physical Sciences Laboratory), **David D. Turner**, **Ulrich Löhnert**, **Derek J. Posselt**, **Kerstin Ebell**, **Gerald G. Mace**, and **Jennifer M. Comstock**. Published online in *BAMS*, September 2020. For the full, citable article, see **DOI:10.1175/BAMS-D-19-0027.1**.

Science relies on observations to develop theories about nature, and to determine if those theories accurately approximate how nature works. A wide variety of in situ and remote sensing techniques are used to characterize, understand, and quantify properties and processes that occur in the atmosphere and at the surface. Improving our understanding of these processes, and how they interact with each other and the environment, is critically important for advancing numerical weather prediction and climate models.

In response to this recognized need, our field has seen an explosion in the number and diversity of remote sensing instrumentation. We are using advanced active remote sensors such as lidars, radars of various wavelengths, sodars, scintillometers, and global positioning systems. We are also using passive remote sensors like infrared spectrometers, microwave radiometers, and imaging radiometers that operate at wavelengths from the visible to the infrared, and beyond. All of these instruments are taking advantage of various physical laws, many embodied in the principles of radiative transfer, to gain new insights into processes in the atmosphere.

There is a common thread that holds in most of these observations: we are not actually observing what we want to know. We are trying to extract very specific information from the remote sensing observations that are typically only partially related to the variable of interest. In the atmospheric sciences, we often call this inverse process a retrieval. For all of the remote sensors, atmospheric variables of interest must be derived from the observations within an inversion algorithm, which often requires the use of prior constraints.

Very often, if we have a remotely sensed measurement that has some sensitivity to the atmospheric variable we desire, we can compute the signal that we would observe with our remote sensor using a so-called forward model (i.e., the forward process). In other words, if we knew the atmospheric state, which includes here all variables affecting the observed signal, we could reproduce the measurement. These forward models ideally are based upon first principles, so that we have a fair degree of confidence in their fidelity. However, they are often nonlinear, which makes

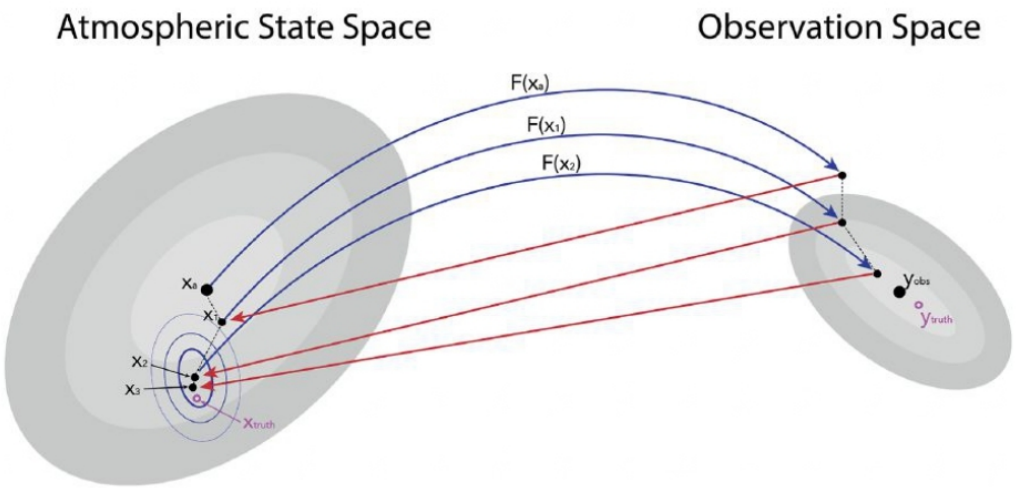

Fig. 1. Optimal estimation principle: The ellipses show the (left) prior state and (right) measurement uncertainty. The iterative process starts with applying the forward operator *F* to the first guess (here $\mathbf{x}_0 = \mathbf{x}_a$). Based on the difference of $F(\mathbf{x}_a)$ to $\mathbf{y}_{obs}$ (which is close to but not equal to an ideal measurement $\mathbf{y}_{truth}$ representing $\mathbf{x}_{truth}$), $\mathbf{x}_1$ is obtained (which requires inverting *F*). This is repeated until the retrieval converges to a solution $\mathbf{x}_3 = \mathbf{x}_{op}$ that is close to the true state $\mathbf{x}_{truth}$.

them difficult, if not impossible, to invert analytically. Thus, the retrieval problem is essentially the development of an algorithm that is used to invert the forward model (*F*) so that we can derive the atmospheric variable that we desire (e.g., humidity, temperature, drop size distribution) from the observation that we have made from our remote sensor (e.g., brightness temperature, radar reflectivity).

Atmospheric research scientist Graeme Stephens provided a classic illustration for a retrieval. Suppose that you desire a description of a dragon, but you observe only the footprints that the dragon makes in the sand. Now, if you already know the dragon, you can pretty easily describe the tracks it might make in the sand (i.e., you can develop a forward model). But if you observe only the tracks in the sand, it will be much more difficult to describe the dragon in any detail. You will likely be able to tell that it was a dragon and not a deer, but there will be aspects that you will be unable to characterize: the dragon's color, if it has wings, etc. A retrieval can combine the observations (large footprints) with prior information (most dragons have wings and the ones making large footprints are green) to get the most likely state (it was a green dragon with wings).

Optimal estimation (OE) is a widely used physical retrieval method that combines measurements, prior information, and the corresponding uncertainties based on Bayes's theorem to find an optimal solution for the atmospheric state with the help of a forward model. In this educational study, we want to stress the importance of properly handling the uncertainties associated with OE. The uncertainty of the retrieval result arises mainly from three sources: first, the prior dataset is critical as a constraint for physical retrievals. Often, however, the data needed to build a well-characterized prior dataset are inadequate or simply unavailable. For example, there are very few observational datasets available that allow us to determine the level-to-level covariance of cloud microphysical properties, which is critical information that is needed for cloud property retrievals, and therefore the community is using other sources such as model simulations. Second, the uncertainties in the forward model assumptions have been neglected for too long and need to be considered. Forward models may be fundamentally incorrect (e.g., applying 1D radiative transfer approaches to situations that are inherently 3D), or may have uncertainties in the model parameters that affect retrieval results. Lastly, perhaps the most obvious source of uncertainty in a retrieval is in the observations themselves. Too often our community assumes that the corresponding measurement covariance matrix is diagonal and that there is no correlation between different measurements within the observational vector. However, it must be stressed that such a detailed error characterization may not lead to a retrieval improvement if other error sources such as measurement biases are not correctly identified before retrieval.

In addition to these three uncertainty sources, users should consider general limitations of OE

such as the assumption of Gaussian uncertainty distributions. Therefore, nonnormally distributed state variables should be normalized to avoid negative impacts on retrieval quality and robustness. Using forward operators that are grossly nonlinear (i.e., are not moderately nonlinear) will also lead to a decrease of accuracy.

Using a series of examples, we show how these uncertainty sources and retrieval assumptions interact and impact the uncertainty of the final retrieved atmospheric state. Readers are strongly encouraged to analyze and modify the examples themselves using supplemental Jupyter Notebooks[1] that can be run online in a web browser. Together with our novel pyOptimalEstimation Python library,[2] this gives the readers all the information needed to get started with their own OE retrieval projects.

[1] Jupyter Notebooks are a web application for creating documents that contain live code, equations, and figures. The supplemental notebooks are available online (https://github.com/maahn/pyOptimalEstimation_examples).

[2] https://github.com/maahn/pyOptimalEstimation

It is important to recognize that OE is just one tool, albeit a powerful one, that can be used to retrieve atmospheric information from remote sensing observations. We stress that more work is needed to accurately characterize the three sources of uncertainties in the future. If the limitations of OE make it inapplicable to a problem, other physical retrieval approaches such as the computationally expensive Markov chain Monte Carlo method are more appropriate for highly non-Gaussian cases or where the forward model is highly nonlinear.

Lastly, even the best retrievals can only be as good as the underlying observations, stressing the need for enhanced instruments that can constrain retrievals better; this could be achieved using new instrument concepts, improved designs, or smarter sensor synergies.

METADATA

BAMS: What would you like readers to learn from this article?

Maximilian Maahn (University of Colorado and NOAA/Physical Sciences Laboratory): *The goal of our article is to lower the entry threshold for developing inverse retrievals and help the readers avoid typical beginner's mistakes. By providing all the required tools and extensive examples, users can hopefully jump-start solving their own problems.*

Dave Turner (NOAA/Global Systems Lab): *Many people in the* BAMS *community use remote sensing retrievals regularly, and probably do not consider that there could be very large uncertainties in the product (e.g., rain rate from the NWS radar network over the CONUS). I hope that this article helps to illustrate how sensitive these could be, especially those in the prior dataset and model.*

BAMS: How did you become interested in the topic of this article?

MM: *During the government shutdown of 2019–20, I was locked out of all my data that were stored on NOAA computers. So I couldn't make any progress with my main projects. Therefore, I thought about what paper that I could write in "one week" and started working on a manuscript about the pyOptimalEstimation library I had developed. When I reached out to Dave [Turner], I learned that he was also working on an Optimal Estimation paper but with a more educational focus. We joined forces and came up with the concept of the current paper. When the shutdown eventually ended, we were all busy with our other projects again and it took about a year finishing the manuscript. However, for me it is still the "one-week paper" even though it took a little longer to finish.*

DT: *I have a background in theoretical mathematics, and I have always been interested in information content. That drew me to Bayesian retrievals, as the framework lends itself well to quantifying the impact of the observations.*

BAMS: What surprised you the most about the work you document in this article?

DT: *I was surprised how often the parametric uncertainties in the forward model dominate the error budget of the retrieved quantity!*

BAMS: What was the biggest challenge you encountered while doing this work?

MM: *The biggest challenge was to make the manuscript easily accessible for grad students without foregoing discussion of important details lost to keeping the manuscript short.*

DT: *It was challenging to take a complex mathematically based subject, which has a lot of power and beauty, and express it in an easy-to-understand way.*

BAMS: What's next?

DT: *The mathematics of retrievals and data assimilation are very similar. I would like to explore employing some techniques used in the data assimilation community to improve retrieval methods.*