JPL Publication 19-2

Uncertainty Quantification for JPL Retrievals

Michael Turmon
Amy Braverman

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

February 2019

This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

© 2019. All rights reserved.

Contents

1 Executive Summary **1**

2 Introduction **2**
2.1 Verification, validation, and prediction 2
2.2 Uncertainty quantification 3
2.3 Inclusive error analysis 5

3 Approaches to Retrieval Quality Assessment **6**
3.1 Non-probabilistic quality indication 6
3.2 Maximum likelihood and optimal estimation 7
3.3 Complete Bayesian retrievals 8
3.4 Computational barriers to operational use 9

4 An Approach to Uncertainty Quantification **9**
4.1 Remote sensing observing systems 10
4.2 Basis and implementation of OSUEs 11
4.3 Relationship to observing system simulation experiments 13
4.4 Strengths and weaknesses of OSUEs 14

5 Tools and Infrastructure **14**
5.1 Sampling the true state 15
5.2 Simulation and downscaling of spatial fields 16
5.3 Accounting for forward model discrepancy 17
5.4 Surrogate models and emulators 18

6 Workshop, Discussion, and Needs **19**
6.1 Correlated errors, including spatial correlations 19
6.2 Validation of standard errors 20
6.3 UQ for formulation 21
6.4 Communication of uncertainty 21

7 Recommendations **21**

A Appendix: Other Types of OSUE **26**
A.1 Optimal estimation 26
A.2 Forward calculation 27

B Appendix: Workshop Presentations **29**

C Appendix: Workshop Attendees **47**

1 EXECUTIVE SUMMARY

This report defines well-accepted notions of verification, validation, and uncertainty quantification (VVUQ), and places them in the context of retrievals of Earth science data. Uncertainty quantification (UQ) is the process of characterizing the uncertainty of a measurement system relative to the true, physical quantity it measures. UQ is needed when measurements are used for scientific inference, assimilated into models with other data, and used for decision-making and risk analysis. Additionally, UQ can affect formulation of new measurement opportunties.

Many measurements come with associated measures of standard error or bias, but these uncertainty measures are often less accurate than they can be. One reason is the operational need to assume crisp values for many parameters that are in fact uncertain, as well as the failure of certain technical assumptions that are needed by operational error-estimation approaches. UQ provides a way to study which assumptions fail to hold, and a tool to improve the measurement. We describe a particular UQ setup that has been successful for OCO-2, and is applicable to many Earth science retrievals.

Besides our own work in UQ, the basis of this report is discussion from a JPL workshop on UQ that was attended by about 25 practitioners, and subsequent interviews with retrieval scientists. We suggest the following improvements to JPL work practices, with the objective of raising the level of JPL UQ practices to the "state of the practice" attained by leading efforts elsewhere.

RP-1 [*Collaboration*] Develop an engagement process that identifies Laboratory needs in UQ and parcels out research tasks to academics, postdoctoral associates, graduate students, and summer students. Engagement would be coordinated by the program office, mission scientists, and UQ points of contact in the line. A research topic list in sec. 7 supplies examples.

RP-2 [*Collaboration*] Develop more robust connections between scientists, retrieval algorithm developers, and line UQ specialists. The core objective is to connect domain scientists to UQ technologists rather than to make scientists experts in UQ.

RP-3 [*Formulation*] Having an articulated plan for UQ, as well as for conventional V&V, should be a requirement for new mission proposals. Define a scale of UQ maturity in the formulation context to accommodate varying needs and infrastructure readiness. Exploit existing forward simulators, OSSE frameworks, and model runs, with modifications as needed. Develop proposal scaffolding for UQ. (sec. 6.3)

RP-4 [*SDS Design*] Explore UQ requirements that impact science data system (SDS) design and implementation. Specific topics include technologies for incorporation of uncer-

tainties throughout data pipelines, establishing standards for reporting uncertainties, and developing extensions for spatial/temporal covariances. (sec. 3.4, 5.2, 6.1)

RP-5 [*Implementation*] Develop a unified infrastructure for systematic UQ that can be applied in the production stream, across different missions or types of retrievals. (sec. 5)

2 INTRODUCTION

Remote sensing uses space or aircraft measurements to find estimates of quantities of interest (QOI) for later use by scientists, Earth system modelers, and decision-makers. As the field has matured from reporting specific measured values to solving the inverse problems relating measured values to the desired physical variables, associating estimates with an uncertainty has become critical. In the context of retrievals of terrestrial state variables, we can distinguish three use cases that drive the need for uncertainties. Scientists interpreting measurements need uncertainties to pose falsifible tests of scientific hypotheses. Earth system modelers need uncertainties to combine measurements; for instance, when measurements are assimilated into state-based models. Finally, decision-makers need uncertainties to understand the likelihood of various outcomes. This report addresses these issues by proposing approaches for quantifying the uncertainty of existing measurement systems. Besides this, the tools used to compute uncertainties of given data products are also useful elsewhere in the mission life cycle. For instance, in formulating new missions, the same approaches that are used to assess measurement uncertainty can identify whether a new measurement technique has enough accuracy to meet mission goals. The error budgets and measurement biases that emerge from uncertainty assessment can also point out opportunities for improved measurements.

2.1 Verification, validation, and prediction

Recovery of uncertainties in complex retrievals is challenging. In standard remote-sensing terminology, the uncertainties of Level 1 data may be assigned directly using known detector characteristics, but Level 2 data typically result from inversions applied to the sensing chain. Uncertainties in the original data will be magnified by the retrieval algorithm, and further increased due to use of ancillary data which may itself have attached or assigned uncertainties. The formalism shown in Figure 1 is helpful in decomposing the elements of this problem [NRC12]. An actually-existing *physical system* — such as the effect of atmospheric composition on sensor radiances — is represented by a *mathematical model*. This model is abstract, and can be thought of as a set of definitions and equations in a textbook. In turn, the equations of the mathematical model are realized by a *computational model* run-

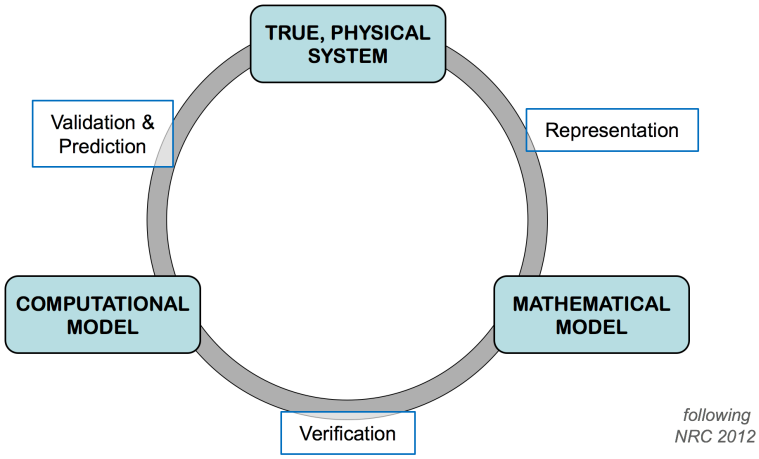


FIGURE 1: *Relationships between models and the verification, validation, and prediction processes.*

ning in a digital computer, with the intent that the computational model shadows the true, physical system under the conditions of interest.

Several processes, also shown, help ensure this intent is realized. *Verification* is the process of ensuring that the computational model faithfully implements the equations of the mathematical model. This is typically a numerical analysis problem that can be addressed over a broad span of realizations of the physical system without close reference to the particular problem at hand. In the retrieval setting, this could correspond to ensuring that radiative transfer codes are numerically accurate as forward models. *Validation* is the process of conducting experiments to ensure that the computational model is an accurate model of the physical system. Validation experiments are performed under known conditions so that ground truth data are available. *Prediction* is the process of using the computational model to infer the state of the physical system under unknown conditions. In the retrieval setting, validation might be done by making observations near an in situ sensor, and comparing the retrieved value with this ground truth. On the other hand, prediction refers to normal instrument operations, away from the controlled environment of validation experiments. Characterization of measurement uncertainty for prediction is the chief concern of this report.

2.2 Uncertainty quantification

Implicit in the notions of validation and prediction is that some measure exists of an acceptable discrepancy between the physical system and the computational model, and this is where *uncertainty quantification* (UQ) enters: UQ is the process of characterizing the uncertainty of a measurement system relative to the true, physical quantity it measures. Figure 2

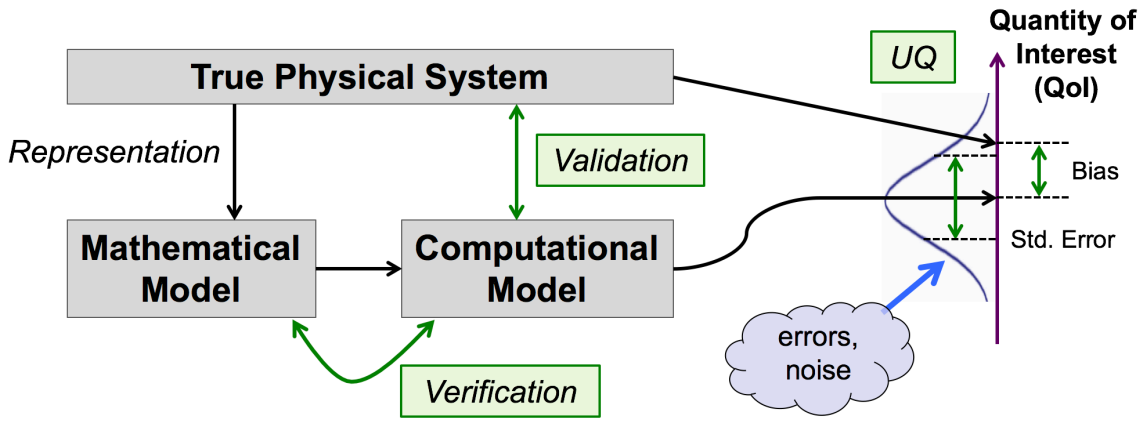


FIGURE 2: *Validation, prediction, and uncertainty quantification, with figures of merit for a QOI.*

recasts the above problem to highlight UQ of a QOI shown at right. Again, the mathematical model abstractly represents the true, physical system, and the computational model implements the math-model equations in computer calculations that have been verified. Validation experiments measure the degree to which the computational model reflects the true, physical system, especially for the most important QOIs. In a validation experiment, a given QOI can be extracted from the physical system and compared to one deduced from the computational model: the results from UQ inform this comparison. And furthermore, at prediction time, when an estimate of the QOI is produced operationally without a ground-truth check, the error assessment from UQ is used to inform downstream users about the accuracy of the measurement. Accuracy can be characterized in terms of bias, shown as an offset of the model from the truth, and standard error, shown as the width of the error bars around the estimated QOI.

Figure 3 shows how these concepts manifest in the retrieval setting. An unobserved true state of nature X passes through a possibly-nonlinear sensing mechanism F, controlled by parameters B. The instrument observes this signature through noise ϵ, yielding an observation Y, corresponding to Level 1 data in sensor units. This observation is then processed by a retrieval algorithm R, controlled by parameters $\hat{B}$, to produce a state estimate $\hat{X}$, corresponding to a Level 2 or higher data product. The parameters may themselves be a result of an estimation procedure that relies on ancillary data with its own error characteristics. In this context, the computational model encompasses the chain above from X to Y, optionally including the retrieval step. A validation experiment would correspond to comparing the state estimate $\hat{X}$ to the true state X, in cases where X is accessible.

This decomposition illustrates the many ways error enters the sensing chain, including:

- Incompleteness of the true state X;
- Limitations of the forward model F or its parameters B in describing the mapping of

the natural state to an observation;

- The stochastic instrument error ϵ;
- Inaccurate values for the ancillary data $\hat{B}$ relative to the true values B;
- Degeneracies in the retrieval algorithm R when given noisy data and inaccurate models.

As with all models of complex systems, it is not obvious which sources of error are dominant, meaning that heritage and experience are important in bounding the scope and fidelity of the model components. Once this has been done, validation experiments provide a way to check the effectiveness of the overall measurement system.

This list of error mechanisms expands upon the terse cloud reading "errors, noise" in Figure 2. Both hint that there are two broad categories of errors: epistemic (due to limitations of knowledge, etymologically descended from the same roots as 'epistemology') and aleatory (random in nature, descended from terms used for dice). Aleatory errors are conceptually simpler — pertaining to things like detector noise or inaccurate parameter values — whereas epistemic errors may not always be well-modeled by probability calculus. For instance, two authorities may differ on the direction of an effect caused by some ancillary variable. The direction is in fact one way or the other, but which is not known. In this case, an average, or a weighting as supplied by conventional probability, is not meaningful. However, such cases of pure epistemic conflict are rare, and a simulation environment that is capable of handling uncertainties due to aleatory errors is often also capable of simulating epistemic effects. For these reasons, as well as for its tractability and analytic power, we use conventional mathematical probability throughout to describe errors and uncertainty.

2.3 Inclusive error analysis

The significance of UQ for the retrieval problems encountered by JPL is that it offers a way to find all-inclusive error bars around measurements, in a way that extends existing verification and validation processes. We attempt to illustrate, in a general setting, the gap in error

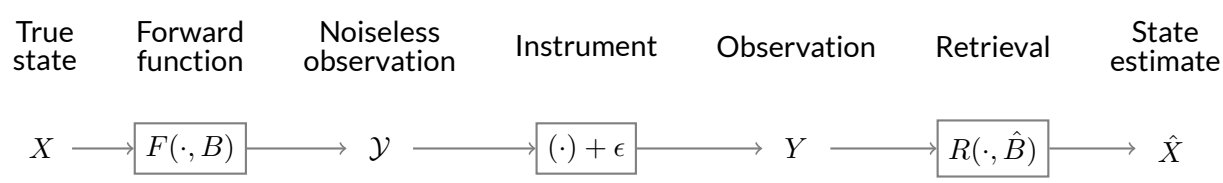


FIGURE 3: *A simplified remote sensing observing system. A state of nature X passes through a forward model F, is contaminated by instrument error ϵ, and is then inverted into a state estimate $\hat{X}$.*

assessment that UQ attempts to close. The observing system uses Y to learn about a QOI X, and the general way to capture this information is the posterior distribution $P(X\,|\,Y)$, as shown on the right of Figure 2. If we fix an observation Y, this distribution, being a function of X, is too unwieldy — it gives the relative probability of any value of X — so we typically summarize it with the mean, $E[X\,|\,Y]$, and the variance, var$(X\,|\,Y)$. We could take $\hat{X}$ to be the former, and the uncertainty to be the latter, and the retrieval problem would be solved. Indeed, well-known results in estimation theory show that these particular choices have many desirable properties.

But, referring to Figure 3, we notice that the retrieval algorithm also uses the ancillary parameters $\hat{B}$ — which are not exactly known. So, what we have actually computed is not $P(X \mid Y)$, but $P(X \mid Y, B)$. In particular, our error bar is seen to be var$(X \mid Y, B)$, and it assumes that B is known exactly. The correct way to adjust the posterior for this is to integrate out against the unknown B:

$$P(X\,|\,Y) = \int_B P(X\,|\,Y, B)\, P(B\,|\,Y) \quad . \tag{1}$$

This averages the posterior, $P(X\,|\,Y, B)$, over all possible values for the unknown parameter B. This integral, though not tractable operationally, always has the effect of broadening the posterior, thereby widening the error bars to account for the extra uncertainty of not knowing B. It could also shift the posterior, indicating a bias. Similar observations hold for the other sources of uncertainty, such as inaccuracies in the forward function F and all other parameter settings within the retrieval algorithm. Each injects its own error, broadening the posterior density. The high-level goal of UQ is thus to account for the error-increasing effects of all hidden assumptions, which can be seen as implicit conditioning. We will revisit this observation below as we consider specific estimation approaches.

3 APPROACHES TO RETRIEVAL QUALITY ASSESSMENT

Every mission is unique and tailors an approach to its own retrieval problem, taking in to account data characteristics, forward model tractability, and computational constraints. Nevertheless, there are unifying principles for all retrievals. Depending on the coarse-level approach adopted, UQ may be formulated in different ways. In this section, we survey different approaches to estimation and to quality assessment, and discuss the capabilities of these approaches to accommodate UQ and the limitations of current practices.

3.1 Non-probabilistic quality indication

Quality flags are often used to characterize error conditions with retrieved data. We are exclusively interested in probabilistic mechanisms in what follows, but some remarks on the

well-known limitations of quality flags may illustrate tradeoffs among different approaches for error characterization. Quality flags are widely *unused* for several reasons in ways that have analogies to the UQ problem.

First, they are not readily interpretable to users. That is, a flag may be set, but outside of an instrument team, the significance of the problem that caused the flag to be set is often not known. Is the problem critical? If a particular quality flag is rarely set, how can I discover whether I need to pay attention to it? The absence of an agreed-upon significance level makes it difficult for downstream users to make intelligent decisions about flagged data. Second, there are quickly too many flags to propagate forward, or to summarize with error conditions that may be functions of multiple flags ("the timing jitter flag is only relevant in nadir pointing mode"). The underlying problem is that there is no calculus to allow meaningful combination of quality bits. These issues (no map between quality indicator and real-world consequence; combinatorial explosion of error conditions) illustrate the power of using probabilistic mechanisms for error characterization where appropriate.

3.2 Maximum likelihood and optimal estimation

In the retrieval setting of Figure 3, an observed Y is used to find an estimate $\hat{X}$ of the true QOI X, in the presence of ancillary parameters B. Maximum likelihood (ML) and optimal estimation (OE) offer solutions to this statistical estimation problem, as well as methods to derive standard errors. Both approaches are used by JPL retrievals; for example, AIRS uses ML, and TES, MLS, and OCO-2 use OE. Maximum-likelihood standard errors are typically derived from either the expected or the observed Fisher information matrix, both of which are related to the curvature of the likelihood function evaluated at the estimate [EH78, Was04, ch. 10]. For remote sensing ("nonlinear regression" in the statistics literature), this curvature reduces to a product of "sensitivity matrices" composed of partial derivatives of F [Smi14, sec. 7.3]. Subject to regularity conditions, these matrices can be related to the standard error of $\hat{X}$.

Optimal estimation standard errors are derived from the posterior distribution of X given Y, $P(X \mid Y)$, which is the product of the same likelihood function used by ML, and a prior on X assumed by OE. The prior has the beneficial effect of regularizing the estimate $\hat{X}$ [Rod00]. Subject to the correctness of the choice of prior distribution, the posterior standard error yields the error of $\hat{X}$. The standard error is computed by combining a sensitivity matrix of derivatives of F (as used by ML) with the prior uncertainty.

Both of these error estimation approaches have limitations. ML relies on a large-sample assumption to extract the estimator standard error [Smi14, sec. 7.2.4], while OE requires an accurate prior. Additionally, the assumption of Gaussian observational errors, which leads to tractable quadratic optimization targets, may not be accurate. These assumptions are

difficult to validate in practice. Another problem with the error estimates produced directly by ML and OE was posed in Section 2.3: the variance injected by ancillary parameters is, in general, not completely accounted for, because some or all of the ancillary parameters B are regarded as fixed, while in practice they contain errors. Furthermore, the assumption of a Gaussian prior on B by OE may be unrealistic for certain parameters. Finally, as mentioned, ML and OE error analysis is typically linearized about the solution, ignoring higher-than-linear terms [CWSM16].

Technical means exist to address many of these limitations, extending the basic ML and OE approaches according to specific measurement needs. UQ provides a mechanism to test these modeling assumptions and serves as an outer control loop around the overall error assessment.

3.3 Complete Bayesian retrievals

Markov chain Monte Carlo (MCMC) techniques allow computation of posterior densities over all parameters, without requiring Gaussian assumptions, and without a linearized error analysis. The resulting posterior can also include the uncertainty contributed by the ancillary data B. Computational capacity limits have generally prohibited operational MCMC computation of full posteriors for Earth science. Experimental, non-operational MCMC has been used in a demonstration capacity for some retrievals in the setting laid out in Figure 3 [TK01, HLL$^+$04, WJYJ13, PM14, BHBM17].

On the other hand, JPL has used MCMC retrievals outside Earth science for the Cosmic Microwave Background spatial field, during the NASA/ESA Planck mission. The retrieved data allows samples to be drawn from the full posterior density, enabling powerful capabilities. For instance, multi-parameter confidence sets recovered from the retrieved data can be intersected with those from other measurements to resolve ambiguities [Pla16, Fig. 15]. Furthermore, these other measurements can be used to perform joint retrieval of concordance parameters, while maintaining the ability to compute confidence intervals on the pooled result [Pla16, sec. 5.6]. In Earth science, this would correspond to forming a a confidence-weighted combination of L2 or L3 data from two disparate instruments (with reference to only the L2/L3 data), a capability that does not now exist.

Evidently, MCMC techniques are poised to make an influential contribution to remote sensing retrievals, in the same way that they have become dominant in other complex estimation problems. (For a sense of the scope of the method, see [Dia09, sec. 6.4].) The key technical element has been the development and implementation of tools for fast sampling of high-dimensional posteriors (hundreds to thousands of dimensions). On the theory side, some key innovations have been development of adaptive, gradient-based, and multi-chain strategies for samplers. This new generation of samplers has shown great ability to solve

high-dimensional inference problems. On the implementation side, there are now several multi-purpose packages (e.g., Stan [CGH+17]) that allow easy experimentation and model development with the above best-practice samplers.

These improved MCMC techniques are a powerful tool, but they are *not a panacea* for solving the UQ problem. This is simply because it is not known, a priori, what effects B should be included in the so-called "full" posterior $P(X \mid Y)$ of (1). That is, MCMC gives the freedom to depart from Gaussian assumptions and linearized analysis of error interactions within the posterior. However, MCMC in itself does not illuminate the issue of which error-causing effects *should* be folded into B. This is where UQ enters: to provide a means of determining experimentally which errors coming from ancillary data or algorithm settings are worth accounting for in the final error analysis. That is, UQ is a mechanism to bound the scope of the error-inducing effects to be included. Similarly, UQ is the process by which the errors produced by any retrieval algorithm (be it MCMC or OE) are evaluated for correctness. This "outer loop" validating the advertised error estimates needs to be present whatever the capabilities of the particular retrieval algorithm used.

3.4 Computational barriers to operational use

At present, off-line (non-operational) uncertainty assessments are generally used, and computational cost is a key reason. For example, as just noted, MCMC retrievals largely remain a case-by-case experiment. The OSUE process, described below, is also not operational. Spatially-extended retrievals will exacerbate the computational problem by introducing more variables into the retrieval, which generally scales worse-than-linear in the number of variables.

One interpretation of this is that we need to introduce computational resources into the trade-space, and make decisions about whether it continues to be acceptable to make simplifying assumptions that produce poorer retrievals at lower computational cost, versus more accurate models and retrievals at higher cost. This is partly a computational issue, and it has not been well-examined from an engineering or cost/benefit point of view. Systems like HySDS (Hybrid-cloud Science Data System, [HMS+15]) provide the computational infrastructure to enable selected retrievals to be run in the cloud as needed.

4 AN APPROACH TO UNCERTAINTY QUANTIFICATION

We describe a generally-applicable approach to UQ for remote sensing retrievals called an Observing System Uncertainty Experiment (OSUE). Ensemble or Monte Carlo approaches are needed, in the light of the limitations listed above: the OE approach does not take uncertainties beyond the Gaussian prior into account, although it is computationally tractable

insofar as it applies. The full Bayesian retrieval takes more uncertainties into account, but it is not tractable. See also Povey and Grainger [PG15, sec. 2.4].

An OSUE is similar in spirit to an Observing System Simulation Experiment (OSSE): it uses a model-based simulation of Nature to produce ground truth, and then mimics the operation of a sensing and data processing system to produce a state estimate that is compared against the reference values. The approach we describe is consistent with international standards for quantifying and reporting uncertainties on physical measurements; see [TK94, Wor09, Wor08a], and compare especially [Wor08b, Wor11]. In this section, we first introduce notation for observing systems, then provide an overview of the OSUE framework, and close with a discussion of the infrastructure needed for implementation of OSUEs at JPL.

4.1 Remote sensing observing systems

Figure 4, top, shows a schematic diagram of a remote sensing observing system, which has three components. Nature generates a true state vector X_0 and defines a mechanism F_0 for converting this true state to a (noiseless) radiance spectrum. Nature also generates a vector of ancillary parameters, B_0, that influences the radiance but is not part of the state vector. The second element of the observing system is an instrument that observes the radiances $\mathcal{Y}_0 = F_0(X_0, B_0)$, contaminated by additive measurement error, ϵ_0, resulting in an observation, Y_0. Finally, a retrieval algorithm, R_0, which typically depends on side parameters, finds an estimate, $\hat{X}_0$, of the original state vector.

Nature's true state, forward function, and noiseless radiance vector are unknown, but scientific theory provides *a priori* constraints upon them. Remote sensing makes inferences about the true state, either from the observations alone (for least squares retrievals) or by explicitly combining them with prior knowledge (for OE retrievals, [Rod00]). To character-

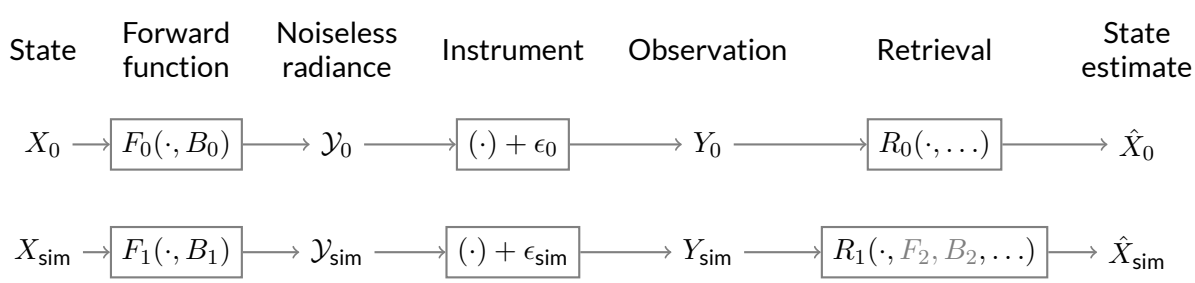


FIGURE 4: *An observing system (top) and OSUE simulation (below). A state vector is mapped through a forward function, F, which depends on ancillary information B, generating a noiseless radiance spectrum, $\mathcal{Y}$. The observing system sees $\mathcal{Y}$ plus noise ϵ. The resulting observation is inverted by a retrieval algorithm, yielding an estimate of the true state vector.*

ize the performance of this observing system, the quantities driving Figure 4 (X_0, B_0, and ϵ_0) can be modeled as draws from probability distributions. Typically, ϵ_0 is taken as independent of X_0, but some elements of B_0 may be dependent on X_0. Having specified the input distributions, the boxed elements correspond to transformations of these inputs, linking the whole collection of random variables, and in particular, X_0 and $\hat{X}_0$. The OSUE proceeds from this idea.

4.2 Basis and implementation of OSUEs

The OSUE (Figure 4, bottom) is a simulation experiment that characterizes transformations of a QOI through a processing chain that estimates that QOI. Since the QOI itself is unknown, the OSUE uses synthetic examples represented by the random variable X_{sim}, and drawn from a representative distribution. As long as the synthetic examples are representative of the real world, and the transformations implemented in the OSUE are accurate depictions of the real processing chain, the statistical properties of the retrieved estimates $\hat{X}_{\text{sim}}$, relative to the synthetic truth, X_{sim}, will mirror those of the actual retrievals $\hat{X}_0$ relative to the QOI, X_0.

Figure 5 highlights the parallel elements of the real observing system and its OSUE model. The left side represents the real world, and the right side is the OSUE simulation. Processing flows from top to bottom. The OSUE studies the relationship between a true QOI (X_0) and its retrieved estimate ($\hat{X}_0$) via the relationship between synthetic QOIs, X_{sim}, and *their* retrieved estimates, $\hat{X}_{\text{sim}}$. The success of the OSUE depends how well the pdf$(X_{\text{sim}}, \hat{X}_{\text{sim}})$ matches that of pdf$(X_0, \hat{X}_0)$, or at least how well the bias and variance of $\hat{X}_{\text{sim}}$ match those of $\hat{X}_0$. The experiment will be successful if we ensure that the elements on the right side of Figure 5 are each accurate representations of the corresponding elements on the left side.

We consider three types of retrieval algorithm: least squares/maximum likelihood, optimal estimation, and forward calculation. The first two are similar in that the retrieved quantity minimizes a cost function, and is thereby found indirectly. MISR is an example of a least-squares retrieval, and OCO-2 is an example of an optimal estimation retrieval. Other retrievals (ECOSTRESS, AVIRIS-NG) are obtained by direct forward calculation of physically-motivated relationships to retrieved, precursor quantities. The three types of retrieval require slightly different versions of the OSUE — we illustrate by describing the least squares OSUE below, and defer the other two approaches to Appendix A.

For the maximum likelihood (ML) retrieval, the OSUE model of Figure 4 applies directly, since the retrieval output, $\hat{X}_{\text{sim}}$, directly estimates the QOI, X_{sim}. Figure 6, depicting the ML OSUE, is therefore the same as the bottom row of Figure 4, except that 1) it indicates, through the use of $\{\cdot\}$ notation, that the simulation experiment uses ensembles of realizations; and 2) it shows the comparison of the matched ensemble members of $\hat{X}_{\text{sim}}$ against

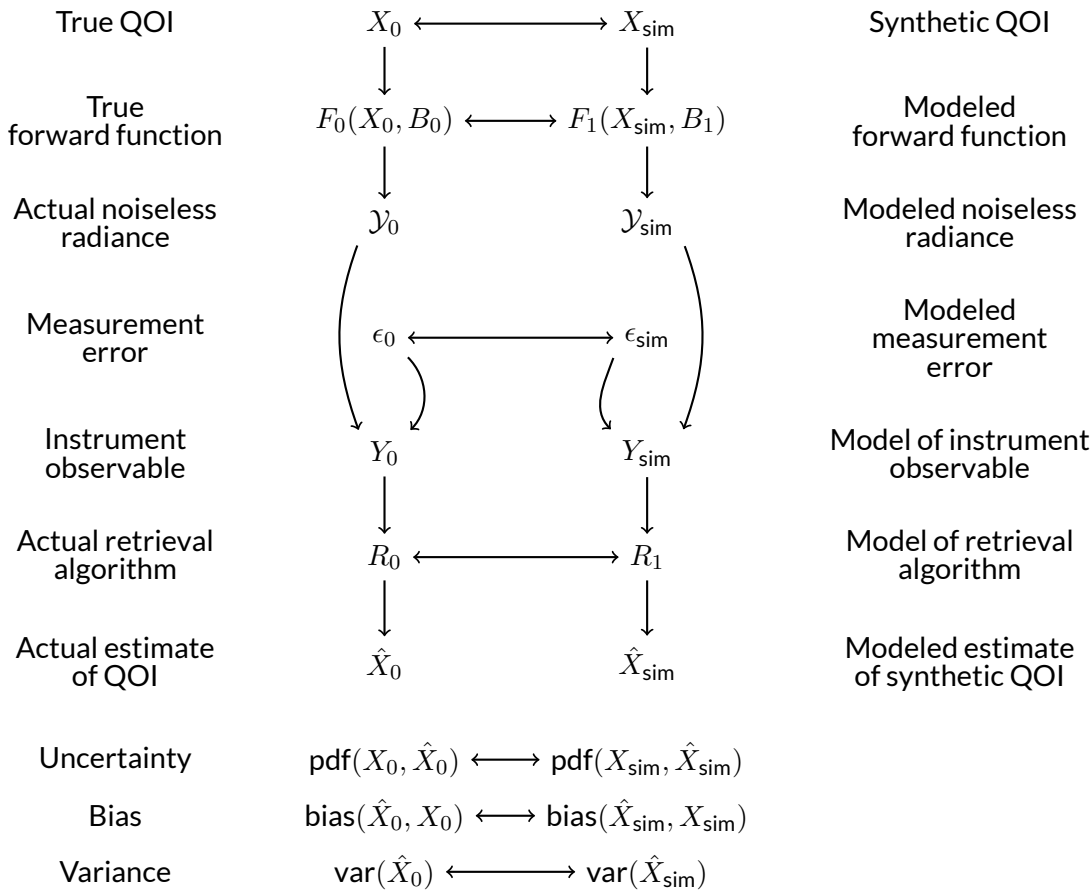


FIGURE 5: *OSUE parallel worlds. The left side represents the real world, while the right represents the OSUE simulation, in which each element faithfully represents its parallel real-world element. Arrows between the parallel tracks show correspondences that must be established by the OSUE design. Correspondences between other elements then hold due to the construction of the experiment.*

X_{sim} to produce estimates of the bias and variance of $\hat{X}_{\text{sim}}$, for instance, the bias of the measurement, $\text{bias}(\hat{X}_{\text{sim}}, X_{\text{sim}})$, is estimated by the OSUE ensemble bias,

$$\widehat{\text{bias}}(\hat{X}_{\text{sim}}, X_{\text{sim}}) = \frac{1}{N} \sum_{i=1}^{N} \hat{X}_{\text{sim}}^{(i)} - X_{\text{sim}}^{(i)} \quad , \tag{2}$$

and likewise for the variance. As noted in sec. 3.2, if certain conditions are satisfied, then standard limit theorems can provide uncertainties of these maximum likelihood retrieval estimates. However, these conditions include linearization of the forward model about the solution, perfect solution of the nonlinear least-squares problem, and a large-sample condition that is hard to satisfy. Finally, some parameters will be assumed as fixed operationally, when in fact only imprecise values are known (sec. 2.3). Encapsulating the retrieval as done in the OSUE allows assessing the impacts of all these factors.

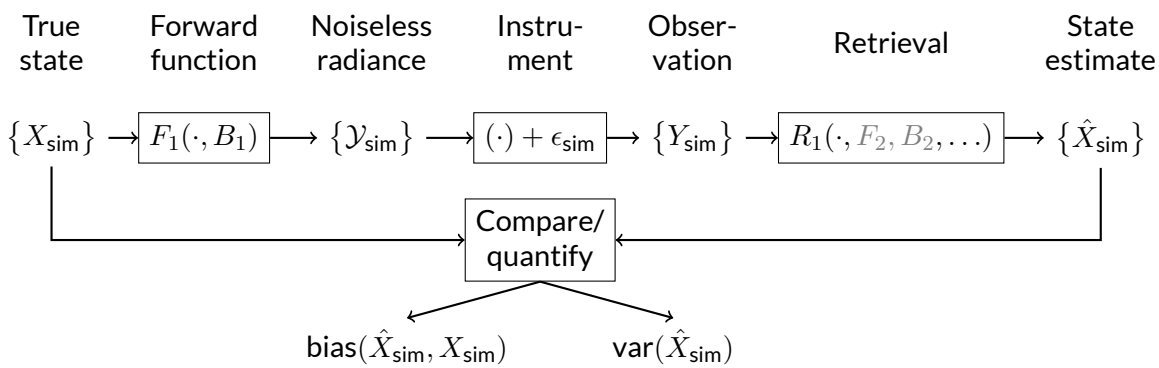


FIGURE 6: *OSUE model of a remote sensing observing system for maximum likelihood retrievals. The ensemble of state estimate vectors is* $\{\hat{X}_{\text{sim}}\}$*, and their true counterparts in the OSUE are corresponding members of the ensemble, input* $\{X_{\text{sim}}\}$*.*

4.3 Relationship to observing system simulation experiments

Observing System Simulation Experiments (OSSEs; [MSE+10]) have long been used to understand how new data sources can improve forecast skill. OSSEs are starting to play a related role in JPL mission formulation as a tool to demonstrate that instruments in formulation will attain the goals expressed in a science traceability matrix (STM). The conventional practice of using small datasets taken under controlled conditions to support traceability is yielding to large-scale computational experiments. By making a quantitative, repeatable connection between mission measurements and science goals, the OSSE can strengthen formulation and enable principled, defensible trade-offs between sensing approach and measurement quality. The OSUE method provides a computational tool to demonstrate that the measurement is adequate, particularly with regard to error margins and sensitivity to environmental assumptions.

OSSE studies for formulation share infrastructure with OSUEs. Indeed, the infrastructure for the conversion of synthetic state vectors to radiances, the instrument model, and application of the retrieval algorithm should be almost identical. The differences between an OSSE and an OSUE are in the way that the synthetic truth is generated, and in how the differences between the retrieved estimates and true states are evaluated.

In an OSSE, the synthetic true state is called a nature run, implying that it comes from a deterministic model of the natural system. Typically, the nature run is computationally expensive, and can only be performed a few times. Parameter settings for those runs are perturbed purposefully, rather than being sampled from a distribution of possible settings. The OSSE therefore is more like a sensitivity study than a randomized Monte Carlo study. Another difference is that the nature run produces a spatial or spatio-temporal field, while the OSUE above produces only individual realizations of the process. Additional research

is needed to develop methods for simulating ensembles of statistical replicates of *spatial fields*, with the proper spatially correlated error structures, from a handful of nature runs.

OSSE infrastructures, which almost always exist for missions and retrieval algorithms in development, can provide the infrastructure needed for such a combined framework. OSSEs will likely have solved the problem of sampling the synthetic true field to create synthetic observed radiances more accurately than would a stand-alone OSUE set up without the knowledge of instrument characteristics available to an OSSE team. The OSUE team can bring principles of experimental design to determining which nature runs can be performed, and potentially how to create additional replicates for the ensemble using statistical methods and emulators. On the evaluation side, the OSUE team can contribute multivariate metrics of agreement and hypothesis testing procedures to determine whether the estimated ensemble of fields agrees with the synthetic true ensemble.

4.4 Strengths and weaknesses of OSUEs

The above OSUE setup has proved effective in UQ for JPL retrievals, and in particular, it has produced many key results on system bias and error for the OE retrieval of OCO-2 [HBC+17]. Some strengths that became apparent are the ability to re-use modules that probably already exist, most obviously, a forward simulator F and a retrieval R, an instrument noise model, and a notion of a nature model to produce X_{sim}. Gradients of F are not needed by the OSUE, except as R may already require them. The Monte Carlo formalism is easy to understand, very general in applicability, and the parallel construction of the OSUE (Figure 5) allows for experimental designs that mimic the way the system is used operationally. One weakness is the computational demands of the Monte Carlo simulations needed for ensemble construction, which may in effect require a fast, reduced-complexity forward model (section 5.4). Another issue is the lack of analytical insight provided by the numerical simulation, which can be partly addressed by comparison with linearized analysis or tailored summary diagnostics of the simulation results.

5 TOOLS AND INFRASTRUCTURE

In this section we discuss some aspects of the implementation of stand-alone OSUEs. These may be regarded as required tools, and their associated infrastructures, for performing uncertainty quantification for geophysical retrievals.

5.1 Sampling the true state

The true distribution of the QOI is unknown, but it is required by the OSUE simulation. The entire OSUE is driven by sampling from the distribution of X_{sim}, so we require that it be a good representation of the marginal distribution of X_0 (see Figure 5), particularly with regard to complete sampling of all observing conditions.

To this end, define a *template set* as any collection of state vectors thought to be representative of the distribution of the QOI, X_0, under circumstances of interest (e.g., "the area off the U.S. Pacific coast during cloudy days in 2016"). The template set can be taken from the output of a physical model, remote sensing or in situ observations, or created by a domain expert. The key is that, if the results of the OSUE are to reflect the performance of the real observing system, the template set must be plausible in type and relative proportion of types. Figure 7 shows a template set derived from a partitioning of OCO-2 state vectors.

To obtain an ensemble of X_{sim} values, note that using the template set itself will lead to ensembles with many repeated values of the same state. Gaussian noise is typically added to the sampled values to broaden the ensemble, but then the covariance matrix of the perturbation must be chosen. Small covariances lead to ensembles concentrated around the templates, while large covariances will cause the centers of mass to be diminished. Gaussian mixture models can provide a flexible way to fit such a distribution to a collection of ensemble values. In fact, re-weighted or non-independent sampling of X_{sim} can produce much faster convergence of ensemble averages such as eq. (2). Such techniques include importance sampling, antithetic variables, and quasi Monte Carlo [Owe13, ch. 8–10].

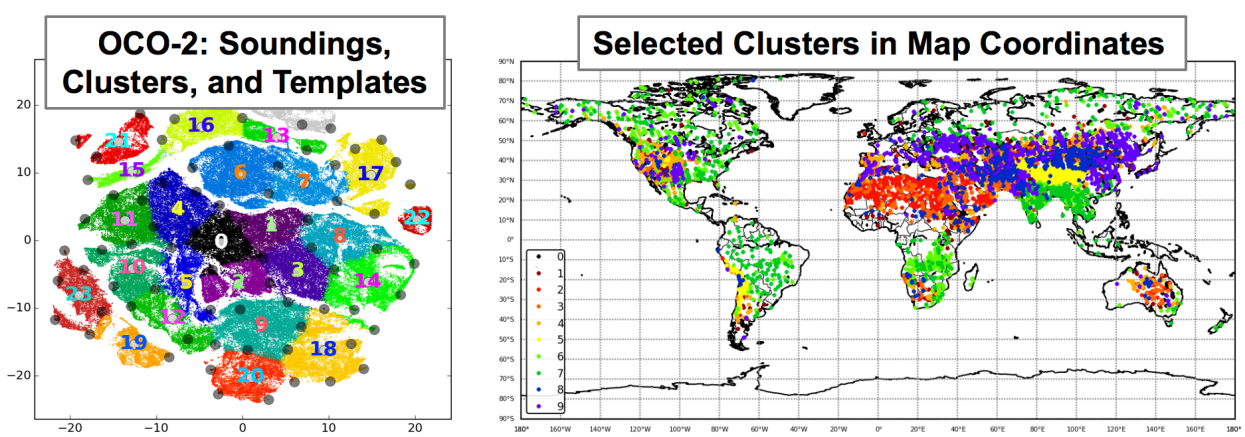


FIGURE 7: *Sampling of OCO soundings: Left panel, two-dimensional projection of 200K soundings (tiny points), partitioned into 25 similar clusters (color-coded), each cluster spanned by representative templates (gray dots). Right panel, projection of selected clusters into map coordinates, revealing some cluster interpretations (green: clouds; red, orange: desert). Result: Lukas Mandrake.*

5.2 Simulation and downscaling of spatial fields

In many applications, footprint-by-footprint standard errors are not enough to characterize measurement uncertainties for later use: error covariances are required. For instance, error covariances are required to compute the error of spatial differences, such as spatial gradients in CO_2 fields as needed for flux modeling, or to compute the standard error of any spatial average. A similar problem can occur in the temporal domain, such as with daily evapotranspiration (ET) cycles as retrieved by ECOSTRESS. Any function of the entire daily ET cycle, such as total daily ET, will have standard error that depends on the correlation of multiple measurements. Error sources like instrument noise will tend to average out over the daily cycle, but systematic errors will reinforce. A full-day UQ will be able to assess this, provided that the temporal covariance structure of the systematic factors is correct. For more on application motivations for such error covariances, see section 6.1; for definiteness, we focus below on the spatial setting.

Estimating spatial covariances requires not just an ensemble of state vectors, but an ensemble of geophysical fields that contain state vectors at all locations. Just as the state vectors must be realistic, the spatial relationships among state vectors in the field must be realistic to achieve accurate uncertainty assessment. In particular, it is not enough to simply add independent noise to the output of a nature run because independent noise does not reflect the spatial covariance. A single nature run is not sufficient to estimate covariance structure without gross simplifying assumptions such as stationarity or isotropy. An ensemble of nature runs would be sufficient, but these are typically expensive to obtain and the perturbations to nature run parameters may be purposeful rather than random.

One path forward is to simulate statistical replicates of one (or a few) model-generated spatial fields using spectral methods (Fourier or wavelet transforms). In this approach, the spatial field of interest is transformed into the spectral domain, and spectral coefficients are partitioned into those associated with scales representing signals and those representing noise. Random perturbations are made to the noise coefficients, and the transform is inverted to yield a new realization of the field [MKBN17]. Research is needed to determine the proper choice of basis, how to set the threshold for defining noise components, and how to generate appropriate random perturbations. Nature runs tend to be at coarser resolution than the footprints of remote sensing observations, but for OSUE purposes, they must be at a comparable resolution. It may be possible to use the spectral approach to simultaneously perturb noise-scale basis function coefficients and downscale the resulting perturbed field.

5.3 Accounting for forward model discrepancy

The forward models used in the retrieval embody the best available knowledge about the system. Of course, the models remain imperfect, and we should include uncertainty due to this *model discrepancy* in the uncertainty estimates. If the best available physical models are already used in the retrievals, then what other models can be used in the OSUE? We discuss three ways to address this problem.

A simple way to deal with model discrepancy is to add an error component to the observed radiances [BO14, Smi14, ch. 12]. As shown in Figure 8, F_1 and B_1 generate radiances, *and* are used in the retrieval R_1. But, a perturbation δ is added to Y_{sim}, compensating for the assumption that the forward model and its parameters are identical to the forward function and its parameters. The perturbation may be Gaussian with a given mean μ_δ and covariance Σ_δ. This simplifies the problem from specifying the mechanistic differences between the forward function and the forward model, to specifying the numerical impact of that difference on the radiances. Information about μ_δ and Σ_δ can be gleaned by examining the differences between actual observed radiances (Y_0 in Figure 5) and radiances computed by the retrieval algorithm during processing ($F_1(\hat{X}_0, B_0)$ in Figure 5).

A related way to deal with model discrepancy is to use the fact that the forward model and the forward model parameters are intimately related, and distinguishing their separate effects may not be possible. That is, consider two forward functions and their associated forward function parameters such that $F_1(X, B_1) = F_2(X, B_2)$. No one can say which forward model is "correct" or which forward model parameters are "correct". We can exploit this lack of identifiability [Smi14] and adopt a single functional form for the forward model and attribute any radiance differences to parameter differences. This recasts the model discrepancy problem as a numerical one, rather than a more abstract comparison of different physics or software implementations.

Finally, in some cases there may exist a progression of forward models ranging from very

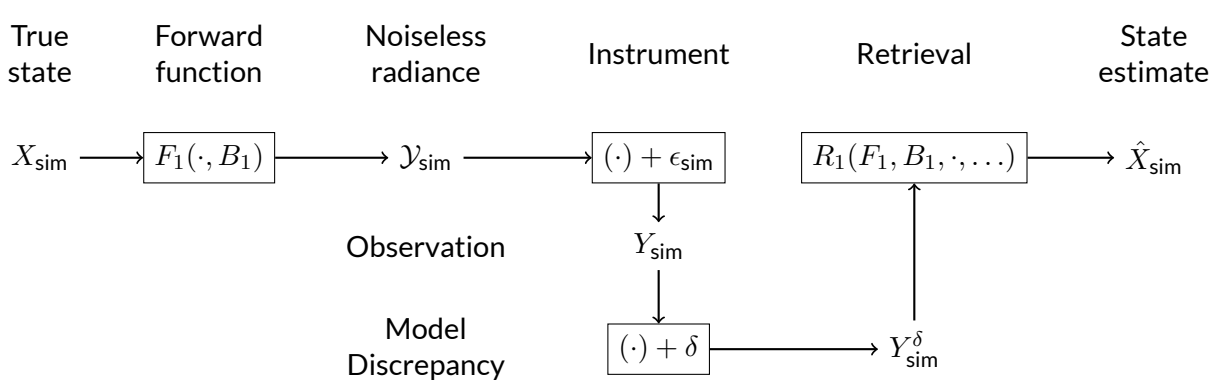


FIGURE 8: *Accounting for model discrepancy by adding a perturbation to the radiance vector.*

realistic to very simple. One can use the more exact models to simulate nature, and the less exact models in the retrieval. For example, if F_0 is nature's true forward function, and F_1 and F_2 are progressively cruder forward models, then one might use $F_1(\cdot, B_1)$ as the forward function in Figure 4 and $F_2(\cdot, B_2)$ in the retrieval. If the relationship between F_1 and F_2 is representative of that between F_0 and F_1, then the experiment would indicate how much uncertainty is added by this model discrepancy.

5.4 Surrogate models and emulators

The computational complexity of the retrieval algorithm may preclude using it in a Monte Carlo experiment, and a simplified version must be used instead. In Figure 5, this possibility is admitted by denoting the actual retrieval algorithm as R_0 and the OSUE retrieval algorithm as R_1. A simplified version of a physical model that can be used in UQ studies is often called a *surrogate model* [Smi14, ch. 13].

For example, a radiative transfer model that accounts for multiple scattering, call it F_1, is relatively slow but more physically accurate than one that does not. A surrogate forward model, say F_2, that implements only single scattering can be embedded in the OSUE retrieval, R_1. This will enable a larger ensemble of radiance vectors to be processed, but it introduces an unwanted source of uncertainty due to the mismatch between $R_0(F_1, \ldots)$ and $R_1(F_2, \ldots)$. The additional uncertainty must be quantified, and the final determination of the uncertainty in the retrieved quantities relative to their respective QOIs must be adjusted. A standard solution is to use tools from Analysis of Variance to quantify the impact of different "treatments" on an outcome variable. Here, the outcome is bias and/or variance, and the treatment is the choice of forward model used in the retrieval. This requires separate, off-line experiments to quantify the contribution to total uncertainty of using a surrogate retrieval rather than the actual retrieval.

If the surrogate model approach is still too slow, an alternative is to use an *emulator*, that is, a function approximator (selected by statistical or machine learning methods) that predicts the output of a physical model as a function of its inputs [Smi14, ch. 13]. Emulators are trained on judiciously selected sets of inputs and corresponding outputs of the physical model (or retrieval). Then, predictions are offered for new inputs of interest. Some emulators (e.g., Gaussian process emulators) have the advantage of automatically producing uncertainty estimates for the prediction, and these can be incorporated into final estimates of bias and variance of the observing system.

6 WORKSHOP, DISCUSSION, AND NEEDS

Table 1 categorizes selected retrievals in operation for JPL measurements as being based on regression, maximum-likelihood (ML), or optimal estimation (OE). It includes microwave, visible/UV/IR, radar, and gravimetric examples. In general, regression-based retrievals directly invert a known physical relationship between the sensed quantity and the desired one. Note that maximum-likelihood in the Gaussian error context reduces to nonlinear least squares estimation, and OE can be viewed as the same least squares estimation but regularized by a prior.

The overview and methodology here were presented to a group of JPL Earth scientists at a "Workshop on VVUQ for Retrievals," held March 27, 2017. About ten of the approximately 25 participants (attendees are listed in the Appendix) offered comments, which are broken down thematically as follows. Where possible, the commenter has been noted.

6.1 Correlated errors, including spatial correlations

Several present (Worden, Landerer, Livesey) raised the question of computation of correlations in errors in the estimates from retrievals. There are two issues: modeling correlated errors in retrieval algorithm inputs, and computing error correlations of multiple retrieved outputs.

The latter received more attention because of its effect on assimilation. Indeed, some current retrievals exhibit multi-variate correlation between separate QOIs, and error correlations across spatial extents (e.g., MLS). Such QOIs with spatially-correlated errors,

$$\hat{X} = X + \epsilon \tag{3}$$

if assimilated into models as if errors are independent, will perturb the state estimate. For example, in the setting where X has a prior mean μ and covariance K_X, and $\hat{X}$ is a noisy observation of X, then the minimum mean square error estimate (MMSE) of X is

$$X_{\text{MMSE}} = \mu + K_X(K_X + K_\epsilon)^{-1}\hat{X} \quad . \tag{4}$$

The reported standard errors are the diagonal entries of the covariance K_ϵ, but the MMSE state estimate depends on the entire matrix. The correlated errors ϵ in (3) will be propagated through (4) without being properly down-weighted. The effects of this problem have been noted in CO2 flux inversion and gravimetric data assimilation.

A consequent question (Yadav) is finding technical means to compactly parameterize spatial or temporal covariances. The reported standard errors, as above, are the diagonal entries of the full error covariance K_ϵ. For most applications, it is not feasible to report the entire covariance. A middle ground could be to report low-rank corrections to a diagonal covariance, but this has not been done operationally for retrievals.

TABLE 1: *Approaches Used for Selected Earth Science Retrievals*

Mission	Target	Approach
TES	O_3	OE
MLS	Trace gases	OE
OCO-2	CO_2, fluorescence (SIF)	OE
Cloudsat	Cloud Water Content (CWC), Cloud Ice Water Content (IWC))	OE
AIRS	H_2O	ML
GRACE	Mass anomaly	OE[1]
MAIA	Particulates and trace gases	OE
Ecostress	Evapotranspiration via brightness temperature	Regression
AVIRIS-NG	Surface reflectance	Regression, OE[2]

[1] The JPL RL05 version of the GRACE OE retrieval initially uses a temporally-unlinked prior followed by a second, temporally-chained solution using the first solution as a prior.

[2] The AVIRIS-NG surface reflectance retrieval is ordinarily done by a radiometric correction (regression), but OE approaches have recently been tested.

6.2 Validation of standard errors

We define validation as the NRC report [NRC12] does: processes for determining biases and standard errors where ground truth data is present. (A related but different question is evaluation of stated standard errors in the more general "prediction" case, i.e., away from ground truth data.) Multiple commenters (including Worden, Boening) asked about formalizing and standardizing processes for validation of standard errors. Claims about the quantitative accuracy of measurement uncertainties are also related to another question, below, on communication of uncertainty.

A key question (Boening, Wiese) is scale mismatch in validation experiments against ground truth (model or in situ) that has a different scale than the measurement. A case in point is GRACE, where observational footprints, and accompanying standard errors, on the domains extending 100s of km may conflict with model data, or in situ data like water wells, on much smaller scales. Similar questions but for temporal scale mismatches arise for atmospheric data (Livesey). In situ data from Fluxnet sites may be available at very fine temporal and spatial scales, and need to be validated against a microwave sounder measurement taken over a longer time window.

It is important to tailor validation experiments, both for measurements as well as stan-

dard errors, to specific science goals (B. Kahn). As pointed out in the NRC report [NRC12, sec. 7.1.2] (Mannucci), no estimation procedure can be considered universally "validated," only validated within a domain of applicability that is covered by the observations used in the assessment.

6.3 UQ for formulation

UQ for formulation is outside the focus of this report, but several present noted that the rigor of UQ in formulation can set the stage for later rigor of retrieval UQ (Mannucci, Gunson). Experience has indicated that projects should discuss UQ needs early on (Gunson), and should plan for the UQ needs of science and operational users (Mannucci). One way to address this is to ensure UQ is present in standard templates used in formulation, such as WBS breakdowns and proposal templates, so it is considered in mission planning (Mannucci).

6.4 Communication of uncertainty

Selected high-risk Earth science domains, like groundwater and sea level rise, would benefit from accurate error assessment of measurements so that risks can be evaluated or outcomes could be priced. In some cases, legal consequences can follow from such determinations (Fisher). Use of measurement uncertainties outside purely scientific contexts would depart from Lab convention (Gunson) and would require a higher level of validation and methodological maturity than is available now.

7 RECOMMENDATIONS

We suggest the following improvements to JPL work practices, with the objective of raising the level of JPL UQ practices to the "state of the practice" attained by leading efforts elsewhere.

RP-1 [*Collaboration*] Develop an engagement process that identifies Laboratory needs in UQ and parcels out research tasks to academics, postdoctoral associates, graduate students, and summer students. Engagement would be coordinated by the program office, mission scientists, and UQ points of contact in the line. A research topic list below supplies examples, but actual tasks will be at the intersection of missions and technical demands. (sec. 5)

RP-2 [*Collaboration*] Develop more robust connections between scientists, retrieval algorithm developers, and line UQ specialists. The core objective is to connect domain

scientists to UQ technologists rather than to make scientists experts in UQ.

RP-3 [*Formulation*] Having an articulated plan for UQ, as well as for conventional V&V, should be a requirement for new mission proposals. Define a scale of UQ maturity in the formulation context to accommodate varying needs and infrastructure readiness. Exploit existing forward simulators, OSSE frameworks, and model runs, with modifications as needed. Develop proposal scaffolding for UQ. (sec. 6.3)

RP-4 [*SDS Design*] Explore UQ requirements that impact science data system (SDS) design and implementation. Specific topics include technologies for incorporation of uncertainties throughout data pipelines, establishing standards for reporting uncertainties, and developing extensions for spatial/temporal covariances. (sec. 3.4, 5.2, 6.1)

RP-5 [*Implementation*] Develop a unified infrastructure for systematic UQ that can be applied in the production stream, across different missions or types of retrievals. (sec. 5)

The following research and technology development topics came up most in the workshops and in interviews with practitioners (sec. 6). These topics are the best opportunities we know of to connect Laboratory needs to research, within and outside JPL, that is advancing the state of the art as described in **RP-1** above.

RT-1 Research means to derive better surrogate models or emulators (e.g., neural networks). Many retrieval algorithms are too slow to support Monte Carlo UQ experiments. (sec. 5.4)

RT-2 Perform research to define and compute what it means for one forward model to approximate another one, for example, in radiance space. Find experimental designs that can quantify model discrepancy for common retrieval types, building trust that the approximate model is adequate. (sec. 5.3)

RT-3 Expand UQ to spatial fields, versus computing errors of single-site retrievals. Connect spatial UQ to science requirements, notably, standard errors of Level 3 regridded data, or accuracy of Level 4 model outputs. (sec. 5.2, 6.1)

RT-4 Identify and pursue an opportunity for use of fully-Bayesian techniques (MCMC) for an operational retrieval, thus accounting automatically for more sources of uncertainty. (sec 3.3)

RT-5 Perform research on advanced sampling strategies, such as quasi-Monte Carlo, stratified sampling, or importance sampling, which have the potential to increase accuracy of UQ estimates by orders of magnitude. (sec. 5.1)

References

[BHBM17] J. Brynjarsdottir, J. Hobbs, A. Braverman, and L. Mandrake. Optimal estimation versus MCMC for CO2 retrieval. *Journal of Agricultural, Biological, and Environmental Statistics*, 2017. In press.

[BO14] J. Brynjarsdottir and A. O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, 30, 2014.

[CGH+17] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

[CWSM16] N Cressie, R Wang, M Smyth, and C E Miller. Statistical bias and variance for the regularized inverse problem: Application to space-based atmospheric CO2 retrievals. *Journal of Geophysical Research-Atmospheres*, 121(10):5526–5537, January 2016.

[Dia09] P. Diaconis. The Markov chain Monte Carlo revolution. *Bull. Amer. Math. Soc.*, 46:179–205, 2009.

[EH78] B. Efron and D. V. Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3):457–487, 1978.

[HBC+17] J. Hobbs, A. Braverman, N. Cressie, R. Granat, and Michael Gunson. Simulation-based uncertainty quantification for estimating atmospheric CO2 from satellite data. *SIAM Jour. Uncert. Quant.*, 5(1):956–985, 2017.

[HLL+04] H. Haario, M. Laine, M. Lehtinen, E. Saksman, and J. Tamminen. Markov chain Monte Carlo methods for high dimensional inversion in remote sensing. *Jour. Royal Statistical Society Series B*, 66(3):591–607, August 2004.

[HMS+15] H. Hua, G. Manipon, M. Starch, L. B. Dang, P. Southam, B. D. Wilson, C. Avis, A. Chang, C. Cheng, M. Smyth, J. L. McDuffie, and P. Ramirez. High-resiliency and auto-scaling of large-scale cloud computing for OCO-2 L2 full physics processing. In *American Geophysical Union Fall Meeting*, 2015. Abstract IN43B-1733.

[MKBN17] P. Ma, E. L. Kang, A. J. Braverman, and H. M. Nguyen. Spatial statistical downscaling for constructing high-resolution nature runs in global observing system simulation experiments. *Technometrics*, 2017. Submitted.

[MSE+10] M. Masutani, T.W. Schlatter, R.M. Errico, A. Stoffelen, E. Andersson, W. Lahoz, J.S. Woollen, G.D. Emmitt, L-P. Riishojgaard, and S.J. Lord. Observing system simulation experiments. In *Data Assimilation, Making Sense of Observations*. Springer, 2010.

[NRC12] NRC Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification. *Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification*. National Academies Press, 2012.

[Owe13] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013. Unpublished, see http://statweb.stanford.edu/~owen/mc/.

[PG15] A. C. Povey and R. G. Grainger. Known and unknown unknowns: uncertainty estimation in satellite remote sensing. *Atmospheric Measurement Techniques*, 8(11):4699–4718, 2015.

[Pla16] Planck Collaboration. Planck 2015 results XIII. Cosmological parameters. *Astronomy & Astrophysics*, 594, Oct 2016.

[PM14] D. J. Posselt and G. G. Mace. MCMC-based assessment of the error characteristics of a surface-based combined radar–passive microwave cloud property retrieval. *Journal of Applied Meteorology and Climatology*, 53(8):2034–2057, 2014.

[Rod00] C. Rodgers. *Inverse Methods for Atmospheric Sounding*. World Scientific, 2000.

[Smi14] R. Smith. *Uncertainty Quantification*. Society for Industrial and Applied Mathematics, 2014.

[TK94] B. N. Taylor and C. E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical report, Physics Laboratory, National Institute for Standards and Technology, 1994.

[TK01] J. Tamminen and E. Kyrölä. Bayesian solution for nonlinear and non-Gaussian inverse problems by Markov chain Monte Carlo method. *Journal of Geophysical Research-Atmospheres*, 106(D13):14377–14390, July 2001.

[Was04] L. Wasserman. *All of Statistics*. Springer, 2004.

[WJYJ13] Y. Wang, X. Jiang, B. Yu, and M. Jiang. A hierarchical bayesian approach for aerosol retrieval using MISR data. *Jour. American Statistical Association*, 108(502):483–493, 2013.

[Wor08a] Working Group 1 of the Joint Committee for Guides in Metrology. *Uncertainty of measurement, Part 3: Guide to the expression of uncertainty in measurement*. ISO/IEC, 2008.

[Wor08b] Working Group 1 of the Joint Committee for Guides in Metrology. *Uncertainty of measurement, Part 3: Guide to the expression of uncertainty in measurement, Supplement 1: Propagation of distributions using a Monte Carlo method*. ISO/IEC, 2008.

[Wor09] Working Group 1 of the Joint Committee for Guides in Metrology. *Uncertainty of measurement, Part 1: Introduction to the expression of uncertainty in measurement*. ISO/IEC, 2009.

[Wor11] Working Group 1 of the Joint Committee for Guides in Metrology. *Uncertainty of measurement, Part 3: Guide to the expression of uncertainty in measurement, Supplement 2: Extension to any number of output quantities*. ISO/IEC, 2011.

A APPENDIX: OTHER TYPES OF OSUE

A.1 Optimal estimation

Optimal estimation (OE; [Rod00]) is based on Bayes' Theorem:

$$P(X\,|\,Y) = \frac{P(Y\,|\,X)}{P(Y)}\,P(X) \propto P(Y\,|\,X)\,P(X), \tag{5}$$

where X is the true state, Y is the observed radiance vector, $P(X)$ and $P(Y)$ are the marginal distributions of X and Y, respectively, and the vertical bar in $P(X \mid Y)$ and $P(Y \mid X)$ indicates conditional probability distributions. Optimal estimation retrievals solve for the mean vector and covariance matrix of the posterior distribution, $P(X\,|\,Y)$: the distribution of the true state, given the observed radiances, under the assumption that the distributions of X, and Y given X, are Gaussian.

It is crucial to appreciate that *the output of an OE retrieval is a distribution*: the Gaussian distribution with mean vector conventionally denoted by $\hat{X}$, and covariance matrix denoted by $\hat{S}$. In particular, the output of the OE retrieval is not a direct estimate of the true state's particular realization at the time and location of the observation — rather, the output is a posterior distribution. Therefore, the goal of the OSUE in this case is to quantify the bias

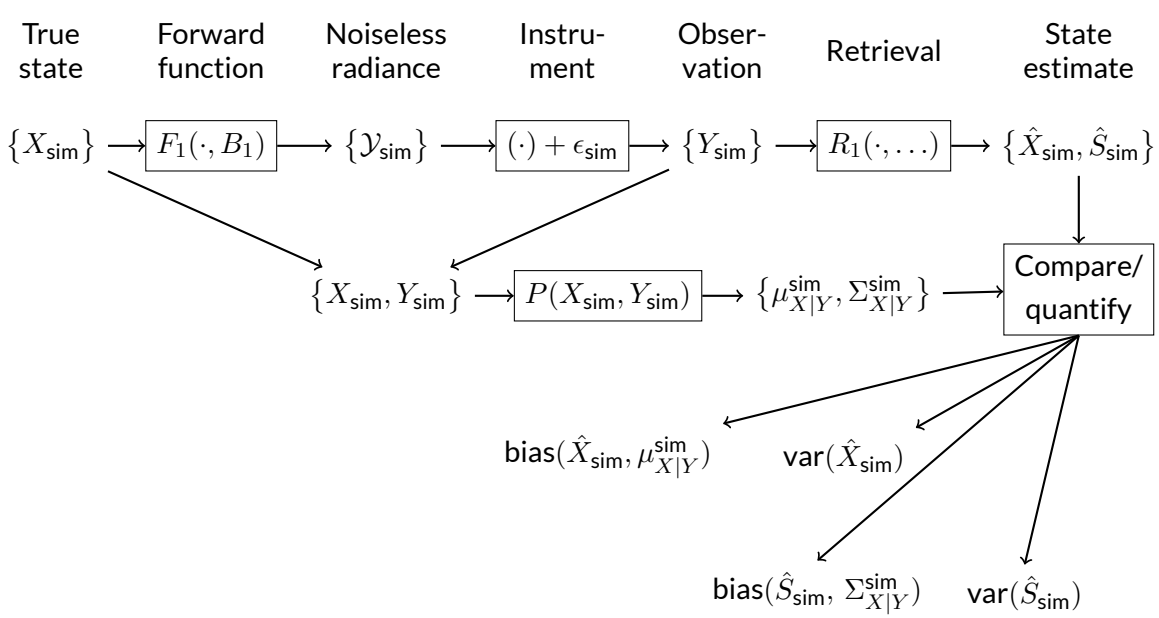


FIGURE 9: *OSUE model of a remote sensing observing system for optimal estimation retrievals. Since the state estimate is a pair of parameters of a distribution, the state estimate must be compared to its true counterpart, the first two moments of the conditional distribution X_{sim} given Y_{sim}: $\mu^{\text{sim}}_{X|Y}$ and $\Sigma^{\text{sim}}_{X|Y}$.*

and variance of the retrieved Gaussian parameter estimates relative to the true parameter estimates. This makes the OSUE in this case more complex than for least squares/maximum likelihood. See Figure 9.

We denote the true parameters of the posterior by $\mu^{\text{sim}}_{X|Y}$ and $\Sigma^{\text{sim}}_{X|Y}$, respectively. These are the quantities estimated by the retrieved posterior mean $\hat{X}$ and the retrieved posterior covariance matrix $\hat{S}$. In the OSUE environment, the true parameters $\mu^{\text{sim}}_{X|Y}$ and $\Sigma^{\text{sim}}_{X|Y}$ can be determined by fitting a probability model to the joint ensemble, $\{X_{\text{sim}}, Y_{\text{sim}}\}$. This is shown in the second row of Figure 9. We denote the resulting probability model $P(X_{\text{sim}}, Y_{\text{sim}})$. From this joint distribution, the conditional (posterior) distribution $P(X_{\text{sim}} \mid Y_{\text{sim}})$ can be computed. Finally, $\mu^{\text{sim}}_{X|Y}$ and $\Sigma^{\text{sim}}_{X|Y}$ can be derived from $P(X_{\text{sim}} \,|\, Y_{\text{sim}})$. It is then these quantities ($\hat{X}_{\text{sim}}$ and $\mu^{\text{sim}}_{X|Y}$, and the covariances) that are compared during UQ.

A.2 Forward calculation

In some cases, estimates of QOIs are obtained by explicit forward calculations rather than an indirect minimization process. See figure 10. Here, the retrieval G_2 is simply the functional inverse of a forward model, or a mapping of observed values onto an empirically-established functional input-output relationship. Computed standard errors in this case would typically be given by mapping expected measurement standard errors through the same functional relationship G_2 according to standard rules of error propagation. The top row of the figure emphasizes that the retrieved values $\hat{W}_{\text{sim}}$ may not include any or all of the system state X_{sim}.

Despite this difference, the UQ approach here is not that different from the ML case. To drive the comparison, a function G_1 is needed that derives the portion of the state corresponding to the QOI from the full system state X_{sim}; the result is W_{sim}. Then, the bias between the OSUE ground truth ensemble, $\{W_{\text{sim}}\}$ and the estimated ensemble can be readily computed, as can the standard error of the OSUE estimated ensemble.

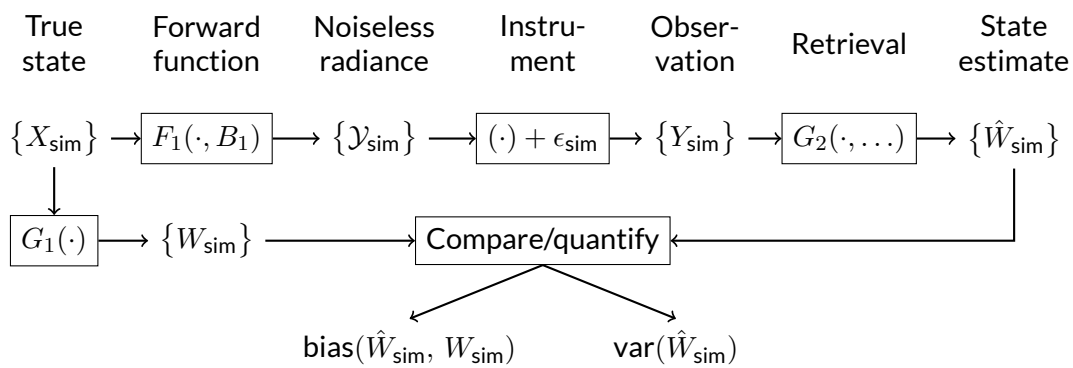


FIGURE 10: *OSUE model of an observing system for estimates based on forward calculation. The ensemble of estimated vectors is* $\{\hat{W}_{\text{sim}}\}$*, and they may not correspond directly to the system state. Their ground-truty counterparts in the OSUE are the corresponding members of the input ensemble* $\{W_{\text{sim}}\}$*, each of which is derived from the corresponding* X_{sim}*.*

B APPENDIX: WORKSHOP PRESENTATIONS

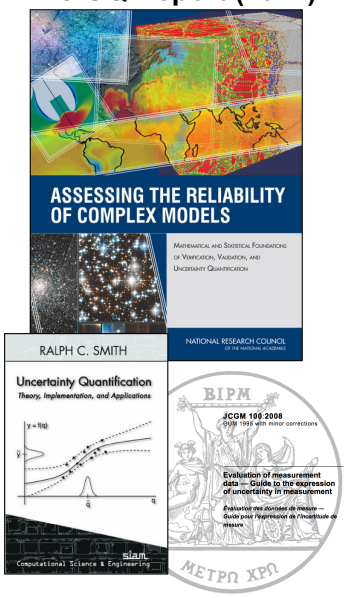
Uncertainty Quantification Overview

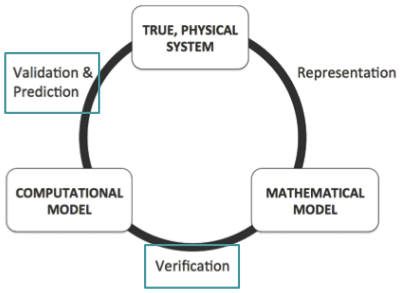
Michael Turmon
contributions from:
Amy Braverman and Jon Hobbs
March 27, 2017

Jet Propulsion Laboratory
California Institute of Technology

UQ Is Situated Within the VVUQ Process

NRC UQ Report (2012)





from NRC 2012

"Verification determines how well the computational model solves the math-model equations,

Validation determines how well the model represents the true physical system, and

Uncertainty quantification (UQ) plays important roles in validation and prediction."

Jet Propulsion Laboratory
California Institute of Technology

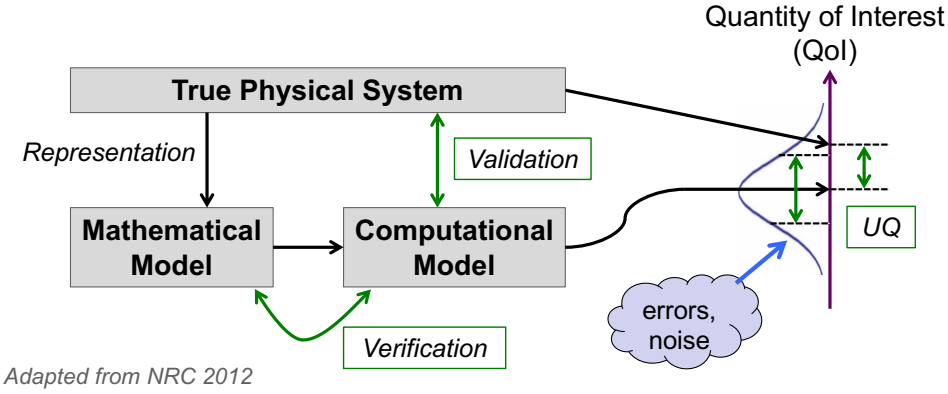
VVUQ Conceptual Relationships

Physical system shadowed by a computational model

Random noise and model errors disrupt estimated QoI

Existing VV processes characterize model fidelity

QoI disruptions are statistical and motivate UQ



Adapted from NRC 2012

Validation, Prediction, and UQ

Jet Propulsion Laboratory
California Institute of Technology

Uncertainty Quantification overlaps *Validation*

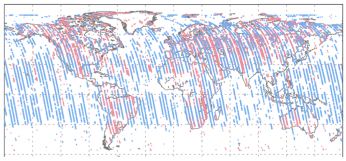
- UQ results should be used in validation processes (NRC, p. 11)

But UQ is especially important for *Prediction* (retrieval)

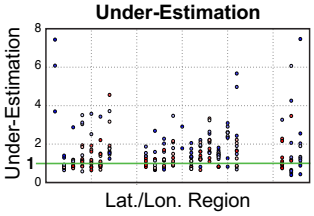
- Previously-controlled variables within validation experiments take on more diverse values (NRC, p.12, 67ff)

OCO-2 Validations

OCO-2 Predictions



OCO-2 Standard Error Under-Estimation



- OCO-2: Validation at ~20 TCCON sites versus global prediction
  - Unknowns can multiply standard errors relative to OE values
- UQ specifically targets parameters critical for measurement re-use: biases, standard errors, and/or confidence intervals

Map: Wunch et al., 2016, AMT. Result: Hobbs et al., 2016, SIAM UQ Conf. + p.c.

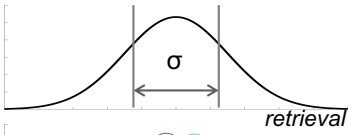
Move Information Left of the Bar

Jet Propulsion Laboratory
California Institute of Technology

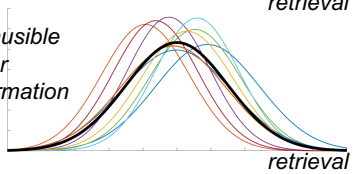
We desire to compute

$$\text{var}(\text{retrieval} \mid \text{obs. data})$$

$$P(\text{retrieval} \mid \text{obs. data})$$



Many plausible values for side information



In fact, we are examining

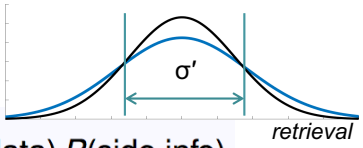
$$P(\text{retrieval} \mid \text{obs. data, side info})$$

Bayesian would recover these extra parameters:

$$P(\text{retrieval, side info} \mid \text{obs. data})$$

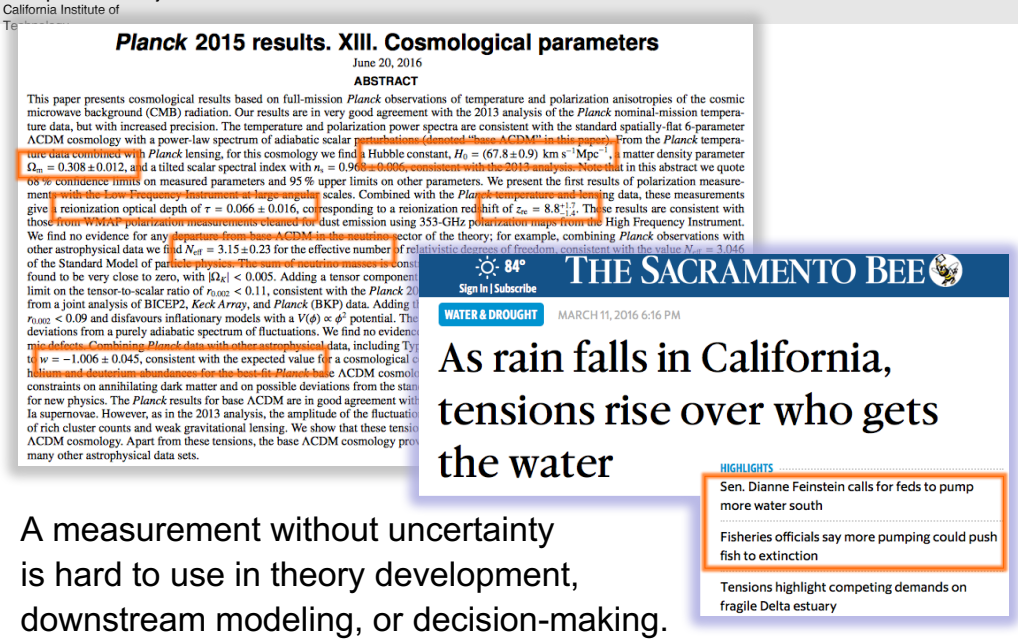
or integrate them out

$$P(\text{retrieval} \mid \text{obs. data}) = \int P(\text{retrieval, side info} \mid \text{obs. data})\, P(\text{side info})$$



Decision-Making Under Uncertainty

Jet Propulsion Laboratory
California Institute of Technology



***Planck* 2015 results. XIII. Cosmological parameters**

June 20, 2016

ABSTRACT

This paper presents cosmological results based on full-mission *Planck* observations of temperature and polarization anisotropies of the cosmic microwave background (CMB) radiation. Our results are in very good agreement with the 2013 analysis of the *Planck* nominal-mission temperature data, but with increased precision. The temperature and polarization power spectra are consistent with the standard spatially-flat 6-parameter ΛCDM cosmology with a power-law spectrum of adiabatic scalar perturbations (denoted "base ΛCDM" in this paper). From the *Planck* temperature data combined with *Planck* lensing, for this cosmology we find a Hubble constant, $H_0 = (67.8 \pm 0.9)$ km s^{-1}Mpc^{-1}, a matter density parameter $\Omega_m = 0.308 \pm 0.012$, and a tilted scalar spectral index with $n_s = 0.968 \pm 0.006$, consistent with the 2013 analysis. Note that in this abstract we quote 68 % confidence limits on measured parameters and 95 % upper limits on other parameters. We present the first results of polarization measurements with the Low Frequency Instrument at large angular scales. Combined with the *Planck* temperature and lensing data, these measurements give a reionization optical depth of $\tau = 0.066 \pm 0.016$, corresponding to a reionization redshift of $z_{re} = 8.8^{+1.7}_{-1.4}$. These results are consistent with those from WMAP polarization measurements cleaned for dust emission using 353-GHz polarization maps from the High Frequency Instrument. We find no evidence for any departure from base ΛCDM in the neutrino sector of the theory; for example, combining *Planck* observations with other astrophysical data we find $N_{eff} = 3.15 \pm 0.23$ for the effective number of relativistic degrees of freedom, consistent with the value $N_{eff} = 3.046$ of the Standard Model of particle physics. The sum of neutrino masses is const found to be very close to zero, with $|\Omega_K| < 0.005$. Adding a tensor component limit on the tensor-to-scalar ratio of $r_{0.002} < 0.11$, consistent with the *Planck* 20 from a joint analysis of BICEP2, *Keck Array*, and *Planck* (BKP) data. Adding t $r_{0.002} < 0.09$ and disfavours inflationary models with a $V(\phi) \propto \phi^2$ potential. The deviations from a purely adiabatic spectrum of fluctuations. We find no evidenc mic defects. Combining *Planck* data with other astrophysical data, including Typ to $w = -1.006 \pm 0.045$, consistent with the expected value for a cosmological c helium and deuterium abundances for the best-fit *Planck* base ΛCDM cosmolo constraints on annihilating dark matter and on possible deviations from the stan for new physics. The *Planck* results for base ΛCDM are in good agreement wit Ia supernovae. However, as in the 2013 analysis, the amplitude of the fluctuatio of rich cluster counts and weak gravitational lensing. We show that these tensio ΛCDM cosmology. Apart from these tensions, the base ΛCDM cosmology prov many other astrophysical data sets.

84° Sign In | Subscribe

THE SACRAMENTO BEE

WATER & DROUGHT MARCH 11, 2016 6:16 PM

As rain falls in California, tensions rise over who gets the water

HIGHLIGHTS

Sen. Dianne Feinstein calls for feds to pump more water south

Fisheries officials say more pumping could push fish to extinction

Tensions highlight competing demands on fragile Delta estuary

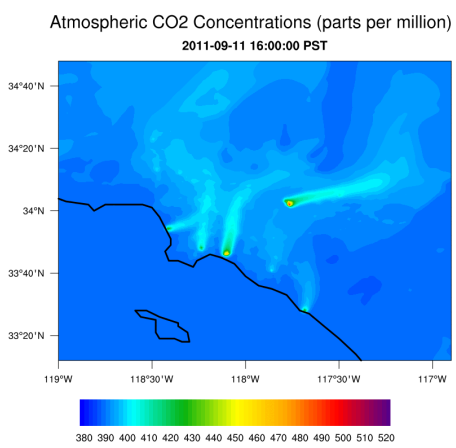
A measurement without uncertainty is hard to use in theory development, downstream modeling, or decision-making.

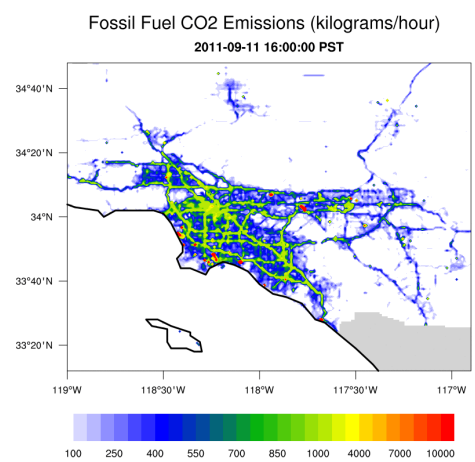
Retrievals vs. Assimilation

Jet Propulsion Laboratory
California Institute of Technology

- UQ for Retrievals can proceed footprint-by-footprint
- UQ for Assimilation uses an entire spatial field
  - UQ requires hard-to-assess spatial/temporal covariance
  - Propagating state uncertainty through time is challenging
- This workshop focuses on retrievals

Results: Vineet Yadav

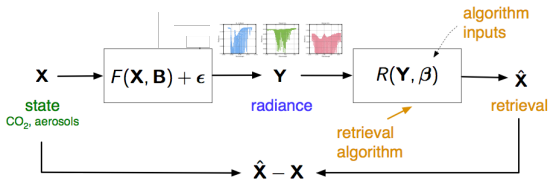




UQ Toolkit: Surrogate Models & Emulators

Jet Propulsion Laboratory
California Institute of Technology

Assume a forward model



Surrogate models
- Mimic a retrieval with lightened computational load
- Enables Monte Carlo studies

Related model-replacement strategies
- Reduced-order models (spectral or other decompositions)
- Spectral methods tested for WSWM hydrology UQ
- Emulators, typically using kriging methodology

The Bayesian toolkit
- MCMC and its variants for Bayesian error bars in non-OE settings
- Heavily used in Planck (7x) and tested in OCO-2

Diagram and formulation: Hobbs, Braverman

UQ Toolkit: Sampling

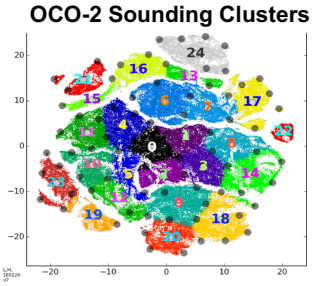
Jet Propulsion Laboratory
California Institute of Technology

NRC report distinguishes between *validation* and *prediction*
- Prediction: find uncertainty under novel conditions, e.g., away from validation sites
- Full posteriors often not computable at operational scale

Partition into *clusters* where posterior variance behaves similarly to a central *template*

Also serves experimental design for uncertainty validation

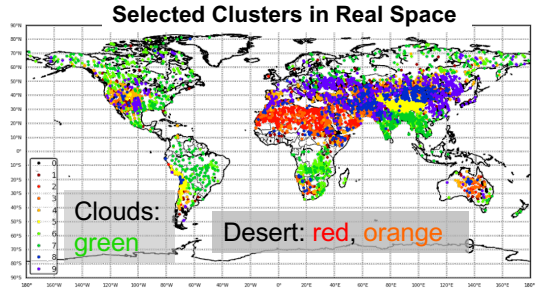
OCO-2 Sounding Clusters



Point: Sounding
Color: Cluster
Dot: Template

Result:
Lukas Mandrake

Selected Clusters in Real Space

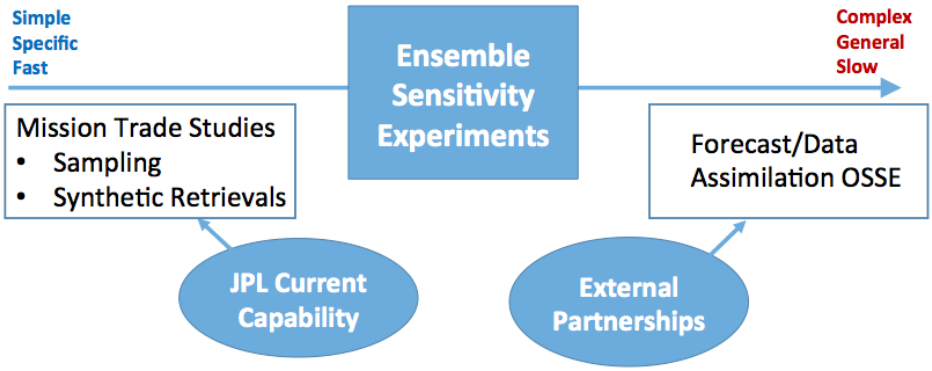


Jet Propulsion Laboratory
California Institute of Technology

Relationship to OSSE

OSSE: Observing System Simulation Experiment

- Require ensembles of underlying conditions
- Forward models (retrieval or event outcome)
- Large-scale Monte Carlo experiments
- Measure and reduce uncertainty of some QoI



Concept and Diagram: Derek Posselt

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

An Uncertainty Quantification Framework for Remote Sensing Retrievals

Amy Braverman

March 27, 2017

1

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Introduction

- Fundamentally, retrievals are complex, algorithmic estimates of QoI's.
- Estimates are subject to uncertainty: the number we compute is not necessarily the "truth".
- Uncertainty = the probability that the estimate is within a specified distance of the true value of the QoI.
- Requires that we quantify the probability distribution of the estimate.

2

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Outline

- Model of an observing system.
- Retrievals as estimators.
- Quantifying uncertainty in retrieval estimates.
- Observing System Uncertainty Experiment (OSUE).
- Example case: OCO-2.

3

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Model of an observing system

True state	Forward function	Radiance	Observation	Retrieval	State estimate

$$\mathcal{X} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\mathcal{X}}$$

$\mathcal{B} =$ other quantities on which the forward function depends.

$\mathbf{B} =$ best proxy for $\mathcal{B}$.

4

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Model of an observing system

True state	Forward function	Radiance	Observation	Retrieval	State estimate

$$\mathcal{X} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\mathcal{X}}$$

- Uncertainty propagation: uncertainty in $\mathbf{Y}$ transformed by R to uncertainty in $\hat{\mathcal{X}}$.
- Additional uncertain quantities injected along the way.
- Uncertainty quantification: uncertainty in $\hat{\mathcal{X}}$ *as an estimate of* $\mathcal{X}$.

$P(\|\hat{\mathcal{X}} - \mathcal{X}\| \leq c) = 0.95$, for example.

5

National Aeronautics and Space Administration

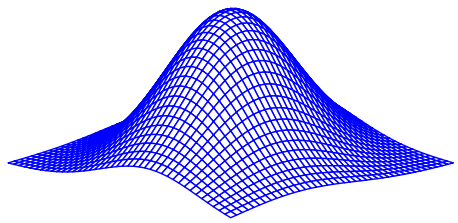
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Retrieval as estimator

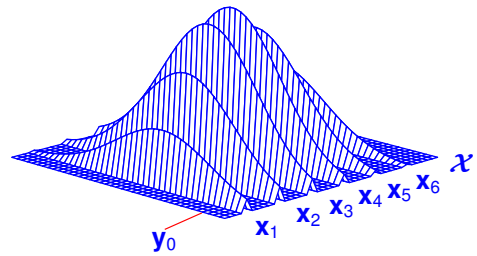
True state | Forward function | Radiance | Observation | Retrieval | State estimate

$$\mathcal{X} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\mathcal{X}}$$

$P(\mathbf{Y}, \mathcal{X})$



$P(\mathbf{Y}, \mathcal{X} = \mathbf{x})$



6

National Aeronautics and Space Administration

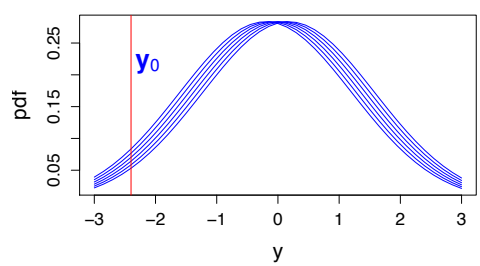
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Retrieval as estimator

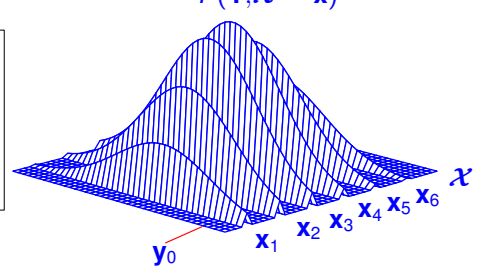
True state | Forward function | Radiance | Observation | Retrieval | State estimate

$$\mathcal{X} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\mathcal{X}}$$

$P(\mathbf{Y} | \mathcal{X} = \mathbf{x})$



$P(\mathbf{Y}, \mathcal{X} = \mathbf{x})$



Maximum likelihood estimate (MLE)

7

National Aeronautics and Space Administration

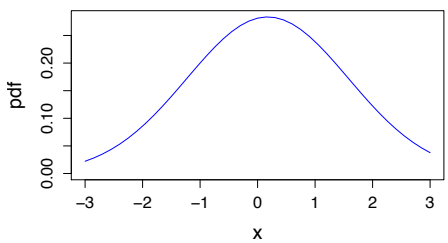
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Retrieval as estimator

True state | Forward function | Radiance | Observation | Retrieval | State estimate

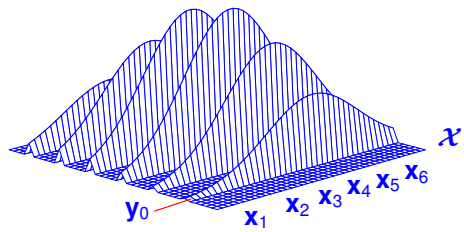
$$\mathcal{X} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\mathcal{X}}$$

$P(\mathcal{X}|\mathbf{Y} = \mathbf{y}_0)$



Bayes estimate (OE)

$P(\mathcal{X}, \mathbf{Y} = \mathbf{y}_0)$



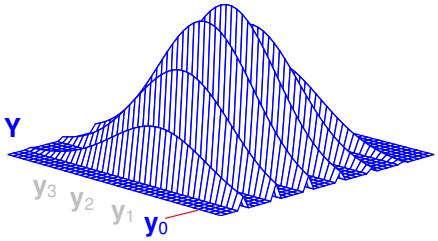
8

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Retrieval as estimator

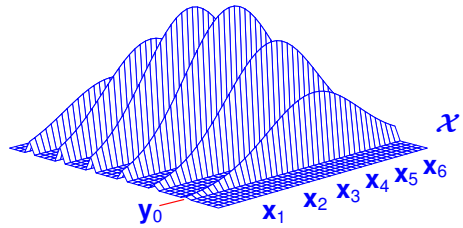
$P(\mathbf{Y}|\mathcal{X} = \mathbf{x})P(\mathcal{X} = \mathbf{x})$



MLE: choose $\hat{\mathcal{X}} = \mathbf{x}$ such that $P(\mathbf{Y} = \mathbf{y}_0|\mathcal{X} = \mathbf{x})$ is maximized.

How *stable* is the MLE? Sample alternative values of $\mathbf{Y}$, construct distribution of $\hat{\mathcal{X}}$'s

$P(\mathcal{X}|\mathbf{Y} = \mathbf{y})P(\mathbf{Y} = \mathbf{y})$



OE: choose $\hat{\mathcal{X}}$ = the first two moments of $P(\mathcal{X}|\mathbf{Y} = \mathbf{y}_0)$.

How *stable* is the OE estimate? Sample alternative values of $\mathbf{Y}$, construct distribution of $\hat{\mathcal{X}}$'s

9

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Retrieval as estimator

A taxonomy of retrieval algorithms:

- MLE (least-squares): AIRS, MISR, SMAP(?)
- OE (Bayes/regularized least-squares): OCO-2, TES, MLS, GRACE, CloudSAT
- Machine learning/statistical: IASI
- Direct forward computation: SWOT(?)

10

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Quantifying retrieval uncertainty

True state	Forward function	Radiance	Observation	Retrieval	State estimate

$$\boldsymbol{\mathcal{X}} \longrightarrow \boxed{\mathcal{F}(\cdot, \mathcal{B})} \longrightarrow \boxed{\mathcal{Y} + \epsilon} \longrightarrow \mathbf{Y} \longrightarrow \boxed{R(\cdot, F, \ldots, \mathbf{B})} \longrightarrow \hat{\boldsymbol{\mathcal{X}}}$$

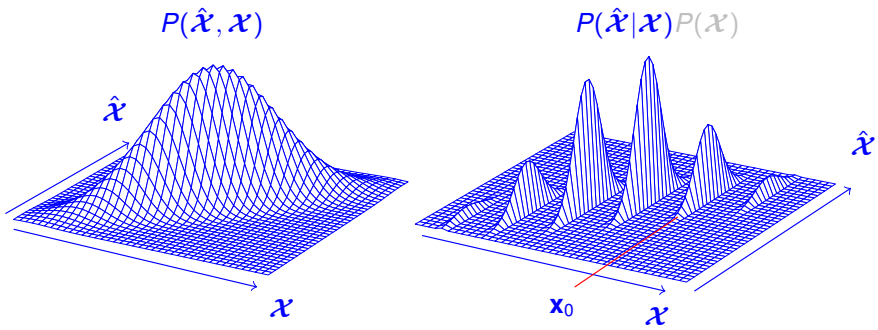
- The stability of the retrieval estimate can be assessed by simulating an ensemble of **Y**'s and calculating the variance of the ensemble of $\hat{\boldsymbol{\mathcal{X}}}$ values (uncertainty propagation).
- But uncertainty is more than variance alone.
- The uncertainty in $\hat{\boldsymbol{\mathcal{X}}}$ as an estimate of $\boldsymbol{\mathcal{X}}$ is fully described by the joint probability distribution of $\hat{\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{\mathcal{X}}$.
- Summarize the relationship between $\hat{\boldsymbol{\mathcal{X}}}$ and $\boldsymbol{\mathcal{X}}$ with bias and variance.

11

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Quantifying retrieval uncertainty



- Bias: is the mean of the distribution $P(\hat{\mathcal{X}}|\mathcal{X} = \mathbf{x}_0)$ equal to $\mathbf{x}_0$ for all possible $\mathbf{x}_0$?
- Variance: what is the variance of $P(\hat{\mathcal{X}}|\mathcal{X} = \mathbf{x}_0)$ for all possible $\mathbf{x}_0$?
- Quantify bias and variance by averaging over possible values of $\mathcal{X}$.

12

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Quantifying retrieval uncertainty

Once we have the joint distribution of the true state and its estimate, $P(\hat{\mathcal{X}}, \mathcal{X})$ (or a summary of it in the form of bias and variance of the estimator), we can make statements like

$$P(\|\hat{\mathcal{X}} - \mathcal{X}\| \leq c) = 0.95,$$

or at least give bounds.

Our approach: simulate $P(\hat{\mathcal{X}}, \mathcal{X})$ with a Monte Carlo experiment– an Observing System Uncertainty Experiment (OSUE).

13

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OSUE

Observing System Uncertainty Experiment (OSUE):

1. Obtain an ensemble of plausible values of $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{X}}_1, \ldots, \boldsymbol{\mathcal{X}}_N$.
2. Simulate corresponding radiances.
3. Process through retrieval algorithm to obtain $\hat{\boldsymbol{\mathcal{X}}}_1, \ldots, \hat{\boldsymbol{\mathcal{X}}}_N$.
4. Empirical ensemble $\{\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}}\}_{n=1}^N$ is a proxy for $P(\hat{\boldsymbol{\mathcal{X}}}, \boldsymbol{\mathcal{X}})$.

Analyze the empirical ensemble to quantify bias and variance representing different geophysical or processing conditions.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OSUE

$$
\begin{array}{ccccccccccc}
\boldsymbol{\mathcal{X}}_1 & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_1 + \epsilon_1} & \longrightarrow & \mathbf{Y}_1 & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_1 \\
\boldsymbol{\mathcal{X}}_2 & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_2 + \epsilon_2} & \longrightarrow & \mathbf{Y}_2 & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_2 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\boldsymbol{\mathcal{X}}_N & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_N + \epsilon_N} & \longrightarrow & \mathbf{Y}_N & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_N
\end{array}
$$

- $\{\boldsymbol{\mathcal{X}}_n\}_{n=1}^N$ are drawn from an ensemble of plausible realizations of the true state.
- $F_0(\cdot, \mathbf{B}_0)$ is a forward function and its parameters, playing the role of nature.
- R is the retrieval function. It's a "black box" in this experiment.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OSUE

$$
\begin{array}{ccccccccccc}
\boldsymbol{\mathcal{X}}_1 & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_1 + \epsilon_1} & \longrightarrow & \mathbf{Y}_1 & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_1 \\
\boldsymbol{\mathcal{X}}_2 & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_2 + \epsilon_2} & \longrightarrow & \mathbf{Y}_2 & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_2 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\boldsymbol{\mathcal{X}}_N & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_N + \epsilon_N} & \longrightarrow & \mathbf{Y}_N & \longrightarrow & \boxed{R} & \longrightarrow & \hat{\boldsymbol{\mathcal{X}}}_N
\end{array}
$$

- For MLE retrievals, straightforward analysis of $\{\hat{\boldsymbol{\mathcal{X}}}_n, \boldsymbol{\mathcal{X}}_n\}_{n=1}^N$.
- For OE retrievals, $\hat{\boldsymbol{\mathcal{X}}}$ is an estimate of the posterior mean and posterior (co)variance (matrix) of the state (vector).
- For OE, the empirical ensemble $\{\boldsymbol{\mathcal{X}}_n, \mathbf{Y}_n\}_{n=1}^N$ approximates $P(\boldsymbol{\mathcal{X}}, \mathbf{Y})$, from which the "true" posterior moments can be deduced.

16

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OCO-2 example

$$
\begin{array}{ccccccccccc}
XCO2_1 & \longleftarrow & \boxed{\boldsymbol{\mathcal{X}}_1} & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_1 + \epsilon_1} & \longrightarrow & \boxed{\mathbf{Y}_1} & \longrightarrow & \boxed{R} \longrightarrow \hat{\mathbf{X}}_1, \hat{\mathbf{S}}_1 \longrightarrow \widehat{XCO2}_1, \hat{S}_1 \\
XCO2_2 & \longleftarrow & \boxed{\boldsymbol{\mathcal{X}}_2} & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_2 + \epsilon_2} & \longrightarrow & \boxed{\mathbf{Y}_2} & \longrightarrow & \boxed{R} \longrightarrow \hat{\mathbf{X}}_2, \hat{\mathbf{S}}_2 \longrightarrow \widehat{XCO2}_2, \hat{S}_2 \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \qquad \vdots \qquad \vdots \\
XCO2_N & \longleftarrow & \boxed{\boldsymbol{\mathcal{X}}_N} & \longrightarrow & \boxed{F_0(\cdot, \mathbf{B}_0)} & \longrightarrow & \boxed{\boldsymbol{\mathcal{Y}}_N + \epsilon_N} & \longrightarrow & \boxed{\mathbf{Y}_N} & \longrightarrow & \boxed{R} \longrightarrow \hat{\mathbf{X}}_N, \hat{\mathbf{S}}_N \longrightarrow \widehat{XCO2}_N, \hat{S}_N
\end{array}
$$

$$\boxed{P(\boldsymbol{\mathcal{X}}, \mathbf{Y})} \propto P(\boldsymbol{\mathcal{X}}|\mathbf{Y}) \Rightarrow (\mu_{\boldsymbol{\mathcal{X}}|\mathbf{Y}}, \Sigma_{\boldsymbol{\mathcal{X}}|\mathbf{Y}}) \Rightarrow (\mu_{\mathrm{XCO2}|\mathbf{Y}}, \sigma^2_{\mathrm{XCO2}|\mathbf{Y}}).$$

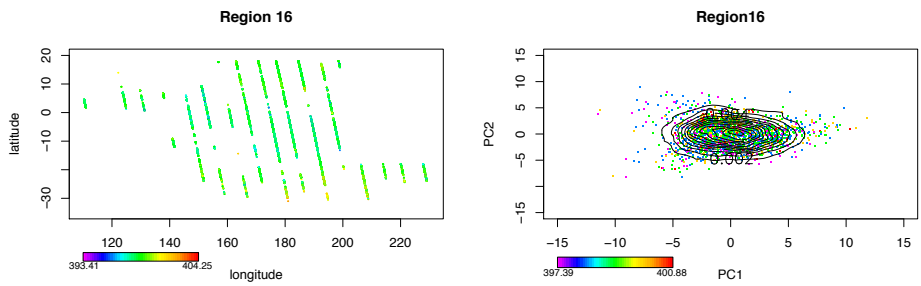
- Ensemble $\{\boldsymbol{\mathcal{X}}_n\}_{n=1}^N$ drawn from a Gaussian mixture model fit to set of representative state vectors.
- Compare distribution of $\{\widehat{XCO2}_n\}_{n=1}^N$ to target value of $\mu_{\mathrm{XCO2}|\mathbf{Y}}$, and the distribution of $\{\hat{S}_n\}_{n=1}^N$ to its target value of $\sigma^2_{\mathrm{XCO2}|\mathbf{Y}}$.

17

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OCO-2 example



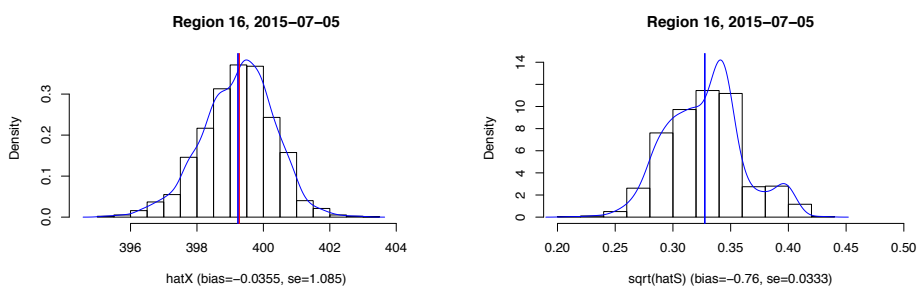
- About 9448 retrievals in Region 16 (south-central Pacific ocean), July 5-11, 2015. State vectors are 45-dimensional. Color scale = retrieved XCO2.
- Sample 2000 times from distribution fit to these data, and use as synthetic truth.
- Upper right panel: two-dimensional projection of the probability density contours of the modeled $P(\mathcal{X}, \mathbf{Y})$.

18

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

OCO-2 example



- Realistic retrieval experiment (almost): all retrieval parameters set to operational values. Forward function = forward model*.
- Left panel: distribution of $\widehat{XCO2}$ relative to $\mu_{\mathrm{XCO2}|\mathbf{Y}}$.
- Right panel: distribution of $\sqrt{\hat{S}}$ relative to $\sigma_{\mathrm{XCO2}|\mathbf{Y}}$.

*It is unrealistic to expect this to be true. We are working on adding a compensating error term.

19

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Discussion

- For this to work for Mission X, Mission X must have:
  - a well-defined state vector that is itself the QoI, or of which the QoI is a known function
  - a realistic forward simulator
  - known measurement error characteristics on the radiances
  - a retrieval algorithm that can ingest synthetic radiances and produce synthetic retrievals fast enough to permit the Monte Carlo simulation
  - information sufficient to create a set of representative states likely to be encountered by the instrument– could be from a physical model, another data source, or even expert judgement

20

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

FAQ

1. How can this be done with out knowing the true forward function $\mathcal{F}$ and its parameters, $\boldsymbol{\mathcal{B}}$?
2. How do you know your ensemble of plausible values covers the space of possible inputs to the retrieval?
3. What if your retrieval algorithm is too slow to process a large ensemble of radiance vectors?
4. How do you use the results to put uncertainties on individual retrievals produced by the operational algorithm?

21

FAQ

5. We are already doing validation- why should we bother with this?
6. Isn't UQ just uncertainty propagation? Why do you care about bias and variance?
7. What if there aren't enough examples of "plausible values"?
8. Doesn't "optimal estimation" already take care of UQ?

C APPENDIX: WORKSHOP ATTENDEES

Unless otherwise noted, affiliations are within the given organizations of the Jet Propulsion Laboratory, California Institute of Technology.

Andrew Bingham, Instrument Software and Science Data Systems Section
Carmen Boening, Earth Science Section
Amy Braverman, Instrument Software and Science Data Systems Section
Sandy Burden, University of Wollongong, Australia
Noel Cressie, University of Wollongong, Australia
Dave Diner, Earth Science Section
Gary Doran, Instrument Software and Science Data Systems Section
Diane Evans, Earth Science and Technology Directorate
Eric Fetzer, Earth Science Section
Josh Fisher, Earth Science Section
Michael Gunson, Earth Science and Technology Directorate
Jonathan Hobbs, Instrument Software and Science Data Systems Section
Simon Hook, Science Division
Brian Kahn, Earth Science Section
Felix Landerer, Earth Science Section
Nathaniel Livesey, Earth Science Section
Anthony Mannucci, Radar Science and Engineering
Derek Posselt, Earth Science Section
Joao Teixeira, Earth Science Section
Michael Turmon, Instrument Software and Science Data Systems Section
Paul von Allmen, Instrument Software and Science Data Systems Section
Duane Waliser, Earth Science and Technology Directorate
David Weiss, Earth Science Section
John Worden, Earth Science Section
Feng Xu, Earth Science Section
Vineet Yadav, Instrument Software and Science Data Systems Section