The Annals of Applied Statistics
2020, Vol. 14, No. 4, 2020–2036
https://doi.org/10.1214/20-AOAS1373

STRUCTURED DISCREPANCY IN BAYESIAN MODEL CALIBRATION FOR CHEMCAM ON THE MARS CURIOSITY ROVER

BY K. SHAM BHAT[1,*], KARY MYERS[1,†], EARL LAWRENCE[1,‡], JAMES COLGAN[2] AND ELIZABETH JUDGE[3]

[1]*Statistical Sciences, Los Alamos National Laboratory,* [*]*bhat9999@lanl.gov;* [†]*kary@lanl.gov;* [‡]*earl@lanl.gov*
[2]*Theoretical Division, Los Alamos National Laboratory, jcolgan@lanl.gov*
[3]*Chemical Diagnostics and Engineering, Los Alamos National Laboratory, bethjudge@lanl.gov*

The Mars rover Curiosity carries an instrument called ChemCam to determine the composition of the soil and rocks via laser-induced breakdown spectroscopy (LIBS). Los Alamos National Laboratory has developed a simulation capability that can predict spectra from ChemCam, but there are major-scale differences between the prediction and observation. This presents a challenge when using Bayesian model calibration to determine the unknown physical parameters that describe the LIBS observations. We present an analysis of LIBS data to support ChemCam based on including a structured discrepancy model in a Bayesian model-calibration scheme. This is both a novel application and an illustration of the importance of setting scientifically informed and constrained discrepancy models within Bayesian model calibration.

1. Introduction. The Mars rover Curiosity was designed to study whether Mars "ever [had] the right environmental conditions to support small life forms" (NASA (2019)). As part of the mission, Curiosity carries an instrument called ChemCam, developed by Los Alamos National Laboratory and L'Institut de Recherche en Astrophysique et Planétologie, to determine the composition of the soil and rocks. ChemCam uses laser-induced breakdown spectroscopy (LIBS) for this task. In LIBS a laser is fired at a target to produce a high-temperature plasma. As the plasma cools, the target emits a spectrum of light over a range of wavelengths that is recorded by a CCD camera. On ChemCam these are captured with three spectrometers covering three different wavelength ranges: ultraviolet (UV), violet (VIO) and visible and near-infrared (VNIR). Figure 1 shows examples of simulated and measured spectra. Each spectrum shows the intensity of light as a function of wavelength. The spectral patterns can be used to identify chemical species and their relative abundances in the target. The presence or absence of certain species and their relative abundances are important clues in answering questions about whether Mars could have ever sustained simple life.

Estimating the chemical composition of soils and rock via LIBS can be difficult. While experts can often easily identify the presence of chemical constituents based on the presence of certain peaks in a spectrum, identifying the relative abundances of the constituents is more difficult due to interactions within the plasma between atoms of the constituents. These interactions, called *matrix effects* (Judge et al. (2016)), can change peak heights in a nonlinear manner. Matrix effects make the disaggregation problem difficult because the spectrum for a target with several chemical species is not a simple linear combination of the spectra for the individual species. Los Alamos National Laboratory has developed a physics simulation code called ATOMIC (Magee et al. (2004)) that can predict spectra in the presence of matrix effects, given inputs such as the chemical species, their relative abundances and other terms that we discuss later. Our long-term goal is to use this simulation capability, along with

Received November 2019; revised July 2020.
Key words and phrases. Bayesian model calibration, simulations, discrepancy modeling, spectroscopy, Mars.

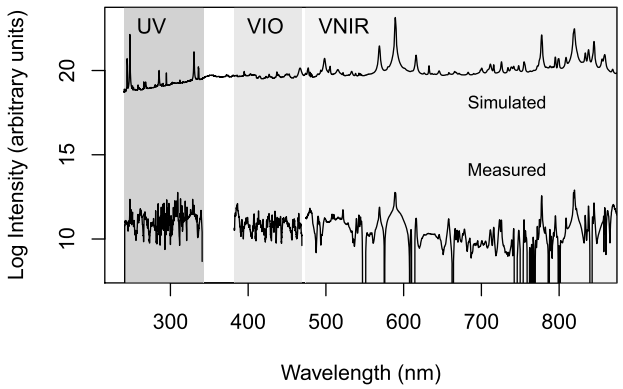

FIG. 1. *Simulated and measured chemical spectra of the compound NaCl over the three ChemCam spectrometers*: *UV, VIO and VNIR. Note that we have plotted these spectra on the log scale to aid our analyses*, *while the spectroscopy community always views spectra on the linear scale. ChemCam measures chemical spectral intensity in photon counts* (*bottom line*), *while ATOMIC computes spectral intensity in power per volume per photon energy per unit solid angle* (*top line*). *This results in clear scaling differences which require quantitative estimation for proper statistical calibration.*

Bayesian model calibration techniques (Kennedy and O'Hagan (2001)), to determine which chemical species are present in a target and in what relative abundances based on measured spectra, thus providing vital clues to answering questions about the history of Mars.

This work focuses on the use of Bayesian model calibration for estimating plasma temperature and density for simpler compounds. This lets us address one particular challenge within this larger problem—accounting for scale differences between simulation and observation. LIBS instruments measure chemical spectra in photon counts as a function of wavelength. In contrast, ATOMIC provides spectra in power per volume per photon energy per unit solid angle. In other words, ATOMIC can predict the shape of the spectrum as a function of inputs. Observed spectra will differ from an ideal simulation by some unknown scaling, as seen in Figure 1. That scaling can depend on many factors, such as laser power, standoff distance and angle and CCD camera properties. Some of these differences can be corrected based on scientific knowledge and testing. We also note that there are other differences beyond the scaling. The measured spectra are noisier, owing to the stochasticity in the measurement process, and there are sharp downward spikes due to very low counts at some wavelengths. The former issue will be handled with a measurement error model, and the latter issue is avoided by concentrating on strong peaks of interest.

In the context of ChemCam, the scientists typically treat this scaling factor as approximately constant within each spectrometer, but it may vary as a function of wavelength. Current efforts (Colgan et al. (2015)) to account for the scaling factor for LIBS spectra do not take advantage of advanced statistical techniques. Thus, they fail to account for potential uncertainty in the scaling difference, make no attempt to quantify the uncertainty in the estimates of the plasma parameters and, ultimately, the chemical composition.

Our scaling difference is a simple instance of a common issue in model calibration: data must be processed because the experiment and simulation do not produce the same output. One example arises in materials science when split Hopkinson pressure bar experiments are used to calibrate flow stress models (Sjue et al. (2020)) but only after the measured strains are converted to stress-strain curves (Gray III (2000)). When done outside of the statistical framework, this process can introduce bias or conceal uncertainty. Our work addresses this by including the difference between theory and observation as a structured discrepancy. This paper presents an analysis of LIBS data that includes estimation of scaling factors as a structured discrepancy within Bayesian model calibration. The structure allows us to strongly

incorporate expert knowledge and avoid problems arising from overly flexible discrepancy specifications (Brynjarsdóttir and O'Hagan (2014)). Additionally, this process provides uncertainty estimates for the plasma parameters as well. In future work we also expect to provide uncertainty for estimates of chemical composition which will provide scientists with a more nuanced understanding of the possible conditions in Mars's past.

Differences between theory and observation of this type are common. In their model calibration work, Kennedy and O'Hagan (2001) initially include two components, one multiplicative and one additive, to account for these differences. One of these is an "unknown regression parameter" that they call ρ that scales the simulation output. This term is typically dropped in later work in this area, apparently for identifiability reasons. One exception is found in Joseph and Melkote (2009).

The other component for modeling differences is an additive discrepancy that they call δ. This term has been the focus of much research. Brynjarsdóttir and O'Hagan (2014) is one of the most important recent references and a key influence on our current work. One of that paper's important conclusions is that "(I)n order to obtain realistic learning about model parameters or to extrapolate outside the range of the observations, it is important not just to incorporate model discrepancy but to model carefully the available prior information about it." This principle guides the construction of our discrepancy modeling. Plumlee (2017) is another important entry in this literature. This work constructs a prior for the bias term that is orthogonal to the gradient of the computer model, alleviating issues with an overly broad posterior on the calibration parameters. Gu and Wang (2018) present an interesting approach to provide good predictions and model fit both with and without the model discrepancy, by using a scaled Gaussian process discrepancy, which more explicitly models the L_2 loss between the simulator and the experimental data. Finally, we note the interesting work of Marmin and Filippone (2018) in which the challenges presented by the traditional framework are addressed using deep Gaussian processes and variational inference to produce a very scalable approach to Bayesian model calibration.

Outside the world of Bayesian model calibration, many people have thought about statistical modeling that includes additional components beyond those that account for the physical process under consideration and, of particular interest here, those that account for effects of the measuring instrument or process. In the case of Raman spectroscopy, Ray and McCreery (1997) present an approach for determining this instrument effect. The problem also arises in astronomical observations of spectra, as in Van Dyk et al. (2001), Lee et al. (2011) and Meng (2018). Statistical quality control methods have been applied to monitor the stability of instrument effects in chromotagraphy (Stover and Brill (1998)). The impact of the detector or sensor on the measured data has also been considered an important source of variability that must be accommodated in statistical modeling under the names of "machine characteristics and performance" in functional magnetic resonance imaging (Genovese, Noll and Eddy (1997)) and "systematic variation" in electrophoresis imaging (Sellers et al. (2007)). We note that these effects of the measuring instrument or process are often called *instrument response*, but we avoid that term here because "instrument response" in the LIBS community refers to a specific effect that is removed during preprocessing. The LIBS version of instrument response could be included in the framework that we present here, but for now we rely on the scientists' preprocessing.

In this paper we present a novel application of Bayesian model calibration to the problem of laser-induced breakdown spectroscopy. This includes an approach to incorporating scale differences between theory and observation into inference based on computationally intensive forward physics models. We use the Bayesian model calibration framework described by Higdon et al. (2008). We will use a highly-structured discrepancy to capture the systematic scale differences. Because of the variation in the simulation output and measured data, we will

operate on the log scale which also allows us to handle the scale difference using the additive discrepancy described above. We will consider two versions of such a discrepancy—one constant, one linear with wavelength—and later discuss a general approach to this problem. Our work will be developed in the context of laboratory measurements that mimic Martian conditions using a ChemCam instrument like the one deployed on the Curiosity rover.

The rest of this paper is organized as follows. In Section 2 we give an overview of measured LIBS spectra from the ChemCam instrument and modeled spectra from the ATOMIC computer model. In Section 3 we provide some background on Bayesian model calibration and present our approach for including scaling factor estimation into model calibration. Next, we demonstrate the approach using perfect model experiments in Section 4 and show results on measured ChemCam data from a laboratory experiment. Finally, we conclude with some discussion and avenues for future research.

2. Laser-induced breakdown spectroscopy. Here, we give some details on the collection of measured spectra using ChemCam and predictions from the theoretical ATOMIC model. In both the experimental and theoretical cases, the output is dependent on the chemical composition of the material being measured or simulated. In the long term our goal is to automate the identification of the chemical species comprising the target material. For the present our goal is narrower—build a modeling approach that can correctly identify unknown scaling differences between theory and measurement, while also correctly estimating a number of other physics parameters. As such, we will not be including a detailed discussion of how the chemical constituents affect measurements nor will we be varying this important input in the theoretical predictions.

All of our data, both measured and theoretical, will come from five relatively pure compounds: KCl, NaCl, SiO_2, Zn and $CaCl_2$. For both theory and measurement, rather than working with the full spectra across all wavelengths, we will focus on the wavelengths around an expert-identified set of elemental peaks that are relevant for those five compounds. These peaks are used by experts to identify the presence of these compounds in LIBS spectra, and we believe that the remaining wavelengths outside these peaks will be relatively uninformative for our task. The wavelength locations of these important spectral peaks are known to have considerable precision, but observed peak locations can be slightly shifted due to spectral adjustments and the resolution of the detector. For a selected peak at a specified wavelength, we consider a surrounding width that should roughly cover the full width of the peak at half its maximum but no more than one nanometer (nm) at each side. These peaks and widths are indicated by the colors in Figure 2.

2.1. *Measured spectra from the ChemCam instrument.* ChemCam measurements begin by firing a laser at a target, such as soil or rock. The laser ablates a tiny portion of the target and produces a small plasma. This excites atoms and emits photons. The light from these photons is optically collected by the CCD camera. A diffraction grating refracts the light, separating the wavelengths and directing the light to the appropriate spectrometer: UV (wavelengths 240–342 nm), VIO (382–470 nm) and VNIR (474–850 nm). The result is a chemical spectrum which is a set of photon counts as a function of wavelength. Different chemical compositions in the target produce different high-intensity spectral peaks. Another source of peaks is the surrounding atmosphere, an unknown proportion of which couples into the plasma, introducing additional peaks into the spectrum. For instance, the Martian atmosphere is 95% CO_2 at a pressure of seven Torr, and this usually results in a noticeable carbon peak at 247 nm regardless of the composition of the target. More information regarding the details of the collection of the ChemCam experiment may be found in Wiens et al. (2013) and Maurice et al. (2012).

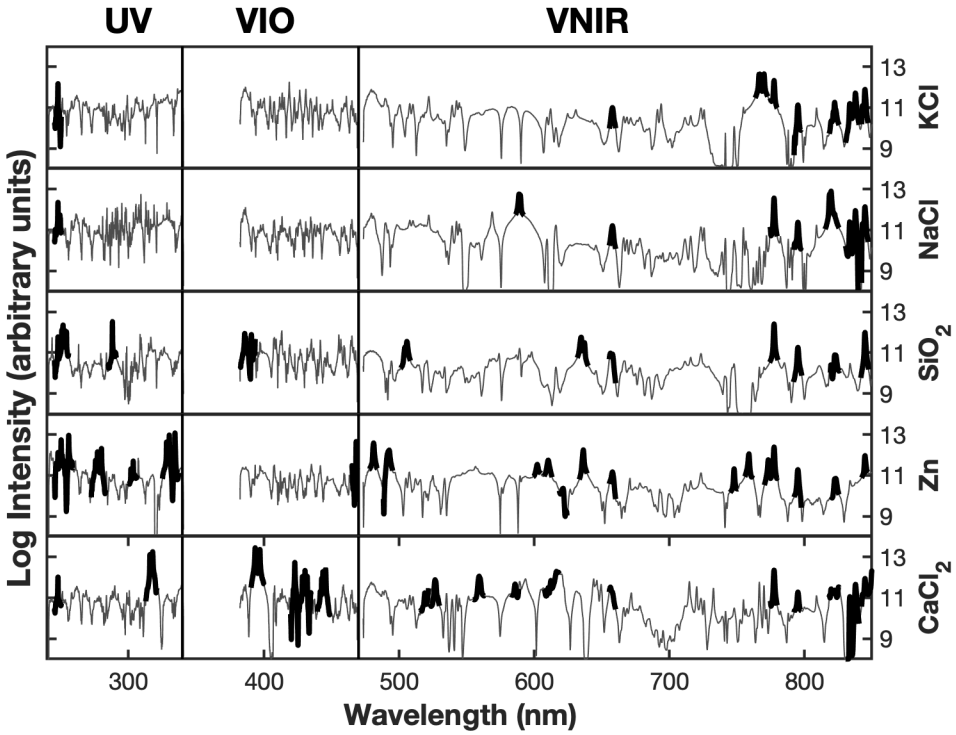

FIG. 2. *Each row shows the observed spectrum for one of our five compounds, computed by averaging the* 75 *laser shots for that compound. The dark regions indicate the expert-identified elemental peaks that were considered to be the most relevant for these five compounds. These selected peaks are overlaid over the full spectrum in each panel. Note that each spectrum includes peaks from the* CO_2 *atmosphere. The vertical lines indicate the wavelength break points* (*at* 342 *nm and* 470 *nm*) *that divide the spectrum across the three instruments*: *UV, VIO and VNIR spectrometers. We will use the selected peaks rather than the full set of wavelengths in each spectrum to estimate the scaling factors and physics parameters in each of the three ChemCam spectrometers.*

Our experimental data, shown in Figure 2, come from laboratory measurements collected with a ChemCam instrument under Martian conditions as follows. For each target, laser shots are repeated 30 times at each of three locations on the target. We discard the first five shots at each location, as they are potentially contaminated by surface dust, and are left with 75 measured spectra for each target. These are postprocessed by the scientists to remove the so-called dark spectra (what the spectrometer records with no light hitting it) and other effects. We examined the shot-to-shot variation in the data and judged it to be small and lacking any structure. Therefore, our observation for each target compound (KCl, NaCl, SiO_2, Zn, and $CaCl_2$) is the average of the 75 postprocessed spectra for that compound.

2.2. *ATOMIC computer model.* The ATOMIC forward model was developed to simulate the emission spectra of chemical compounds using first principles theoretical atomic physics, the details of which can be found in Magee et al. (2004). Briefly, ATOMIC is a general-purpose plasma modeling and kinetics code that has been designed to compute emission (or absorption) spectra from plasmas either in local-thermodynamic equilibrium (LTE) or in non-LTE. The primary inputs are the plasma temperature and density along with a model describing the atomic structure and scattering data of the constituent material(s). ATOMIC receives data (energy levels, transition probabilities, quantum numbers, etc.) from the Los Alamos suite of atomic physics codes (Fontes et al. (2015)). In the simulations discussed in this paper, the results were generated from the CATS code (Cowan (1981)) with modifications made for plasmas generated from LIBS (Colgan et al. (2014)). ATOMIC then uses these

data to compute the average ionization of the plasma (for a given temperature and density) and the resulting emissivity of the plasma. ATOMIC has recently been used to model the emissivity from a number of LIBS plasmas (Colgan et al. (2014, 2015), Judge et al. (2016)).

For each of our five compounds, we run simulations over the space of three parameters:

1. Plasma temperature T, with units electron volts (eV) and range [0.5, 1.5];
2. Mass density ρ, with units of g/cm^3 and range $[-7, -4]$ on the $\log_{10}$ scale;
3. Proportion of target p in the plasma with range [0.1, 0.98]. Because the plasma produced by the laser includes components of the atmosphere near the target of interest, this parameter distinguishes the fraction of the plasma that comes from the target, as opposed to the atmosphere.

This gives a total of 15 parameters (three parameters for each of five compounds). These ranges also indicate the bounds for our uniform prior distributions in the calibration described below to estimate these parameters. The output is a spectrum for the specified compound with intensity as a function of wavelength. The simulation provides intensity as power per volume per photon energy per unit solid angle. ATOMIC predictions account for matrix effects but do not account for any effects arising from the spectrometer itself. Computation time for a single run of ATOMIC for a compound can range from minutes to hours on a high-performance computing system, depending on the chemical complexity of the compound.

3. Calibration with scale differences. As seen in LIBS and other applications, the observations are a scaled version of the output from the physics simulations at some unknown value of the input. In other words, we expect the scaling factor to be multiplicative. However, we will operate on the log scale for all of our modeling. As a result, we will model our scale difference using an additive discrepancy. A number of points argue for the use of the log scale in this problem. First, the rawest data are photon counts, which we might expect to follow a Poisson distribution. Although the data are processed, their Poisson origin suggests a partially multiplicative structure. Further, Michel and Chave (2007) show that repeated LIBS measurements are right-skewed with a few large measurements. Finally, we note that the simulation output itself varies by orders of magnitude over the design. Each of these issues can often be tamed by modeling the data on the log scale. Some exploratory analysis confirms this. As part of that exploratory work, we also attempted a crude version of the analysis on the linear scale and found that the emulator accuracy is reduced and the resulting residuals are right-skewed and heavy-tailed. For this reason we let y^{obs} be the log of a measured spectrum.

3.1. *Computer model calibration.* Bayesian computer model calibration is, by now, a well-studied problem. We will generally follow the approach described in Kennedy and O'Hagan (2001) and extended in Higdon et al. (2008). We will review this briefly. In short, the goal is to find the parameters and systematic biases of a computer simulation that make it best match experimental data.

Assume that an output y, possibly a vector, is observed with measurement error from a true physical system (e.g., a ChemCam measurement). Physical reality can be be approximated by a simulator (e.g., ATOMIC) denoted $\eta(\cdot)$ with parameters θ. This approximation may have systematic biases or discrepancy, denoted δ, which may be a function of the measurement index (e.g., a function of wavelength) or other experimental conditions. Therefore, a general model for the measurement y^{obs}, using the computer simulation $\eta(\cdot)$, is

$$y^{\text{obs}} = \eta(\theta) + \delta + \epsilon, \tag{1}$$

where ϵ captures the measurement error. In our case the observation and simulations consist of the spectra at the selected wavelengths shown in Figure 2. The spectra are concatenated

across compounds. Thus, the observation vector includes all five compounds, as does each run of the simulation. Our parameters are plasma temperature T, mass density ρ and proportion of target p for each of the five compounds, thus

$$\theta = (T_{\mathrm{KCl}}, \rho_{\mathrm{KCl}}, p_{\mathrm{KCl}}, T_{\mathrm{NaCl}}, \rho_{\mathrm{NaCl}}, p_{\mathrm{NaCl}}, T_{\mathrm{SiO_2}}, \rho_{\mathrm{SiO_2}}, p_{\mathrm{SiO_2}}, T_{\mathrm{Zn}}, \rho_{\mathrm{Zn}}, p_{\mathrm{Zn}}, T_{\mathrm{CaCl_2}}, \rho_{\mathrm{CaCl_2}}, p_{\mathrm{CaCl_2}}).$$

The goal in the Bayesian paradigm is to estimate the posterior distribution for θ and δ. Since ATOMIC, like other simulators used in such problems, is computationally expensive, we need to build an emulator, a statistical approximation to the physics simulation. For this work we follow the approach of Higdon et al. (2008), as implemented in the software package GPMSA (Gattiker et al. (2016)). The simulator is run over a design of inputs (described later for our specific case). For each input vector t_i, we observe the simulation result $\eta_i = \eta(t_i)$. These will be used to build the emulator. The emulator is decomposed as

$$\eta(t) = \mu + \sigma \sum_{h=1}^{p_\eta} k_h w_h(t), \tag{2}$$

where μ is the mean vector of the training simulations and σ is a scalar computed as the standard deviation across the output index over all the training runs. The k_h are computed based on a singular value decomposition of the standardized simulations. Let z_i be the ith training simulation standardized with μ and σ. Let $Z = [z_1, \ldots, z_m]$ be the matrix with the standardized training simulations as columns. Compute the singular value decomposition: $Z = USV'$. Let $K = US/\sqrt{m}$. The k_h are the columns of this matrix K which may be truncated to include only those associated with large singular values. We typically choose p_η to account for at least 99% of the total variance in the training simulations. We project Z onto K to get the training weights for each basis vector. The weights $w_h(t)$ for each basis vector are fit using a Gaussian process (GP) over the input design. Hyperpriors for the GP, as well as further details on emulation, are described in Higdon et al. (2008).

The discrepancy model δ is also based on a basis representation,

$$\delta = \sum_{h=1}^{p_\delta} d_h \alpha_h. \tag{3}$$

In Higdon et al. (2008), the basis vectors d_h are normal kernels evenly spaced over the output space (wavelength in our case), and the basis weights α_h have a zero mean Gaussian prior with a marginal precision $\lambda_{\alpha h}$. The number of vectors p_δ, and effectively the spacing between the Gaussian kernels, is chosen based on the desired flexibility of the discrepancy function (fewer discrepancy basis vectors of evenly spaced Gaussian kernels implies longer correlation lengths). If the discrepancy is modeled as a function of experimental conditions, the prior becomes a Gaussian process over this space. The discrepancy can be viewed as a convolution of the kernels d_h and the weights α_h. A weakly informative Gamma(1, 0.001) prior is placed on the $\lambda_{\alpha h}$. Unless the data informs otherwise, $\lambda_{\alpha h}$ will remain at a large value consistent with almost zero discrepancy. Lastly, the measurement error is represented as $\epsilon_i \sim N(0, \frac{1}{\lambda_{yi}} \Sigma_{yi})$, where Σ_{yi} is an $n \times n$ covariance matrix that may be specified by the user.

3.2. *Discrepancy for scale differences.* While the framework of Higdon et al. (2008) provides considerable flexibility, we prefer to constrain any possible discrepancy for both statistical and scientific reasons. Statistically, a flexible discrepancy can cause identifiability problems. A spectrum is composed of a number of peaks. If a discrepancy model over wavelength can make adjustments at the scale of peak widths, then this will be confounded

with physics model adjustments to peak heights. For this reason we need to carefully choose a discrepancy basis that adjusts physics model output on scales that avoid confounding. We also need to ensure that our discrepancy basis is not confounded with parameter effects in the physics code; this is discussed more below.

Scientifically, we expect discrepancy to behave in particular ways, so we should encode that knowledge in a constrained basis. First, as mentioned earlier, the scaling factor is mostly assumed to be constant within each of the spectrometers. However, there are a few other possibilities of scientific interest. In particular, there are well-known LIBS instrument effects at the edges of each of the spectrometers where the sensitivity decreases. These are known well enough to be corrected in postprocessing (Wiens et al. (2012)), but effects may linger. In general, the scaling factor could change as a function of wavelength. For this paper we will consider two simple models: a constant scaling factor for each spectrometer and a scaling factor for each spectrometer that varies linearly with wavelength. The latter is the simplest model that considers variation with wavelength. In Section 5 we discuss extensions of this approach when the modeling is done on the linear scale.

For the constant scaling factor model, we use three basis functions for the discrepancy, one for each spectrometer,

$$\delta_{\text{con}} = d_{\text{UV}}\alpha_{\text{UV}} + d_{\text{VIO}}\alpha_{\text{VIO}} + d_{\text{VNIR}}\alpha_{\text{VNIR}}. \tag{4}$$

Each of the basis vectors d_{UV}, d_{VIO} and d_{VNIR} has length equal to the number of wavelengths in our model. We set the ωth entry $d_{h,\omega} = 1$ for all wavelengths ω in spectrometer h and 0 otherwise. Hence, the basis weights α_k may be interpreted directly as scaling factors.

For the linear scaling factor model, we use six basis vectors for the discrepancy representing the intercept and slope for each of the three spectrometers,

$$\begin{aligned}\delta_{\text{lin}} = {} & d_{\text{UV},0}\alpha_{\text{UV},0} + d_{\text{UV},1}\alpha_{\text{UV},1} \\ & + d_{\text{VIO},0}\alpha_{\text{VIO},0} + d_{\text{VIO},1}\alpha_{\text{VIO},1} \\ & + d_{\text{VNIR},0}\alpha_{\text{VNIR},0} + d_{\text{VNIR},1}\alpha_{\text{VNIR},1}.\end{aligned} \tag{5}$$

Again, all of the basis vectors have length equal to the number of wavelengths in our model. The intercept terms $d_{h,0}$ are identical to the d_h from the constant model. For the slope terms we set the ωth entry $d_{h,1,\omega} = \omega$ when the wavelength ω is in spectrometer h and 0 otherwise. Now, the basis weights describe a linear model for the scaling factor.

Returning to the question of identifiability, we believe that it is especially important to evaluate the parameter effects using the emulator. In our experience in other computer model calibration problems, there are often parameters whose effect is largely in the form of a constant shift or a change in the slope of an output. Therefore, we run the risk that our discrepancy model may mimic the main effect of a physics model parameter. To explore this possibility, we computed a simplified main effect for each parameter in which each physics parameter is varied over its range while all other parameters are fixed at the center of the design range. This is good practice under any circumstance but especially so here. We present the results of this diagnostic study in Section 4.1.

Prior distributions for the ATOMIC model parameters are assumed to be uniform on the parameter bounds given in Section 2.2. We use $p_\eta = 15$ principal component basis vectors to build the emulator. As in Higdon et al. (2008), prior distributions on the discrepancy coefficients are assumed to be Gaussian with zero mean and unknown precision. These precisions are weakly informative Gamma$(1, 0.001)$ priors. We construct the measurement error matrix Σ_{yi} by treating wavelengths independently—based on the assumption of independent Poisson error in the photon counts for each wavelength—and taking the variance over the 75 shots at each wavelength. We estimate the error precision terms λ_{yi} with Gamma priors.

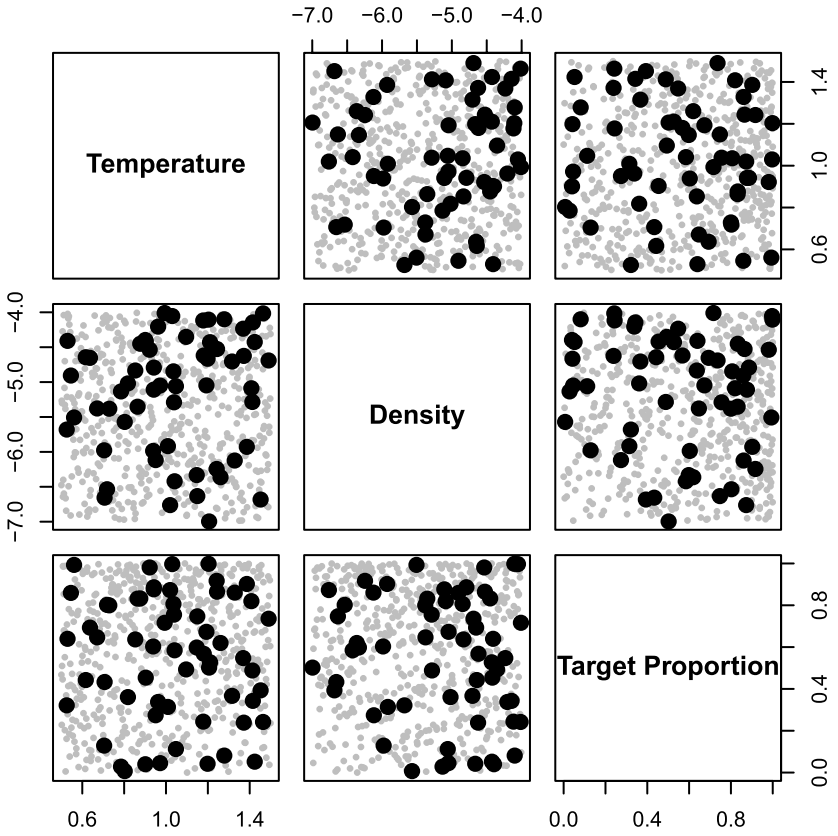

FIG. 3. *Bivariate scatterplot for an $m = 600$ point LHS design in gray, showing a subset of the full design. Here we show the three ATOMIC model parameters T, ρ, and p for one compound, KCl. This is a subset of the full 600-run by 15-parameter design covering the parameters for all five compounds. The black points show a separate set of thirty input settings chosen for emulator testing and evaluation.*

4. Results. Here, we present results and diagnostics from the calibration of ATOMIC spectra to ChemCam spectra to learn about the scaling factors for each of the three spectrometers. For this entire study we use two sets of ATOMIC runs, a subset of which is shown in Figure 3, for emulator construction and evaluation. The gray points are a 600-run Latin hypercube design for training. The black points are a 30-run Latin hybercube design for testing and evaluation. This figure shows the three parameters for KCl, which are a subset of the full 600-run, 15-parameter design covering the parameters for all five compounds. For all estimation we use MATLAB software, called GPMSA (Gattiker et al. (2016)), for emulation and model calibration. GPMSA employs Markov chain Monte Carlo (MCMC) methods to estimate unknown parameters. For the calibration for each of the three spectrometers, we ran the MCMC for 40,000 iterations after a burn-in of 3750 samples using an adaptive step.

Below we first present the results for the emulator diagnostics using cross-validation and main effects. Next, we show the calibration results for a perfect model experiment for both the constant and linear scaling factors, and, last, we give the results for our calibration of the ATOMIC model to the ChemCam LIBS data.

4.1. *Emulator diagnostics.* We begin by evaluating the quality of the emulator using the test set. To do so, we compute $R^2 = 1 - \sigma^2_{\text{res}}/\sigma^2_{\text{raw}}$ where σ^2_{raw} is the variance of the test set around its empirical mean and σ^2_{res} is the variance of the test set residuals, that is, the test set minus the emulator predictions. Computed pointwise at each of the wavelengths that we modeled, the minimum R^2 was 0.9904, suggesting that the emulator does an excellent job.

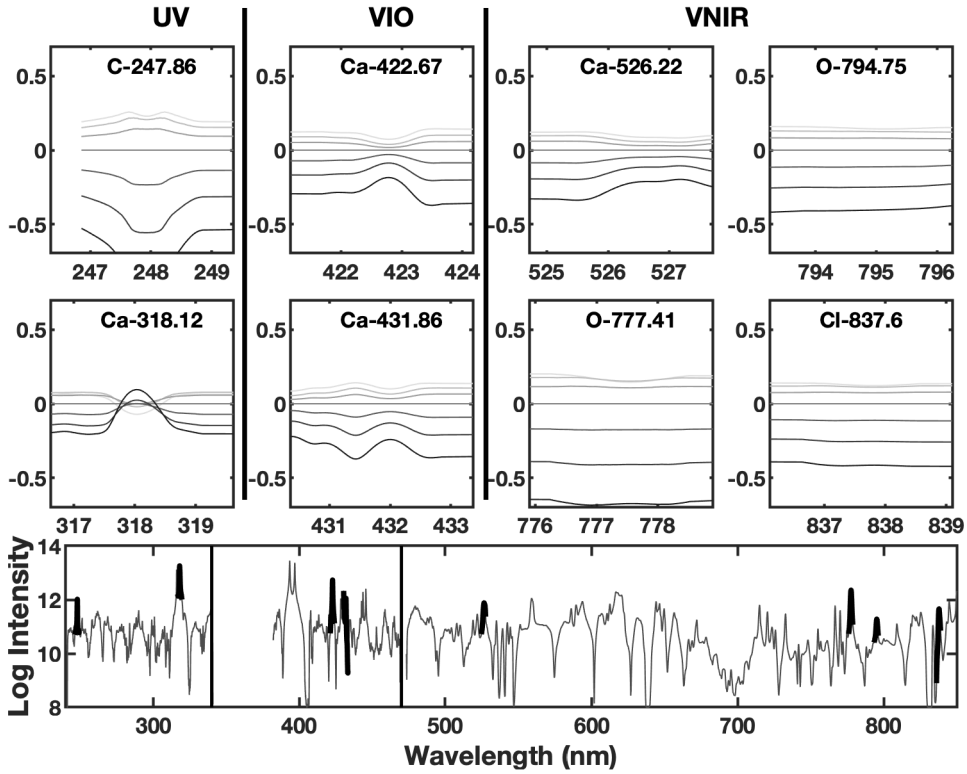

FIG. 4. *Of our* 15 *parameters* (*three input parameters* × *five compounds*), *target proportion* p *for* $CaCl_2$ *appears to be the most likely to be confounded with our discrepancy, and we explore that in this figure. The bottom panel shows the measured spectrum for* $CaCl_2$ *with the eight selected peaks highlighted as in Figure* 2. *The eight panels in the top two rows show the main effect plots for the target proportion of* $CaCl_2$ *for these peaks, presented in increasing order of wavelength and labeled with the element and center wavelength. For each panel, we fix the parameters plasma temperature* T *and density* ρ *at the center of their ranges and vary the target proportion parameter* p *over its domain, with light gray indicating the lower bound and dark gray the upper bound. We subtracted the center prediction from all results. Even here, we expect there to be no problem with identifiability because the carbon peak at* 247.86 *nm has additional structure, and the magnitudes of the other shifts are neither constant across all peaks nor varying with any notable structure, linear or otherwise.*

As discussed earlier, the main effects of the parameters are especially important in this application because of potential identifiability issues with the discrepancy. In particular, we will be concerned if we find parameter effects that cause nearly constant shifts or shifts that are approximately constant and change linearly over length. Our explorations found that the most worrisome parameters in our study are those associated with target proportion p, something we will discuss later in Figure 5. Across our five compounds the effect of p for $CaCl_2$ showed the most potential difficulty, as illustrated in Figure 4. Even here, there seems to be little cause for concern. Although many of the peaks, such as the three oxygen peaks, show roughly constant shifts, other peaks, particularly the carbon peak in the top-left panel, can be used to separate these effects from discrepancy. Even among peaks that show constant shifts, the magnitudes of those shifts are neither constant nor varying with discernible structure across wavelengths.

4.2. *Perfect model experiment*. To demonstrate that computer model calibration approaches can be successful in estimating the ATOMIC model parameters and the scaling factors for chemical spectra output, we perform several perfect model experiments. From the 600 runs that were used to train the emulator, we hold out one run $\theta^{\ddagger}$ and add white noise

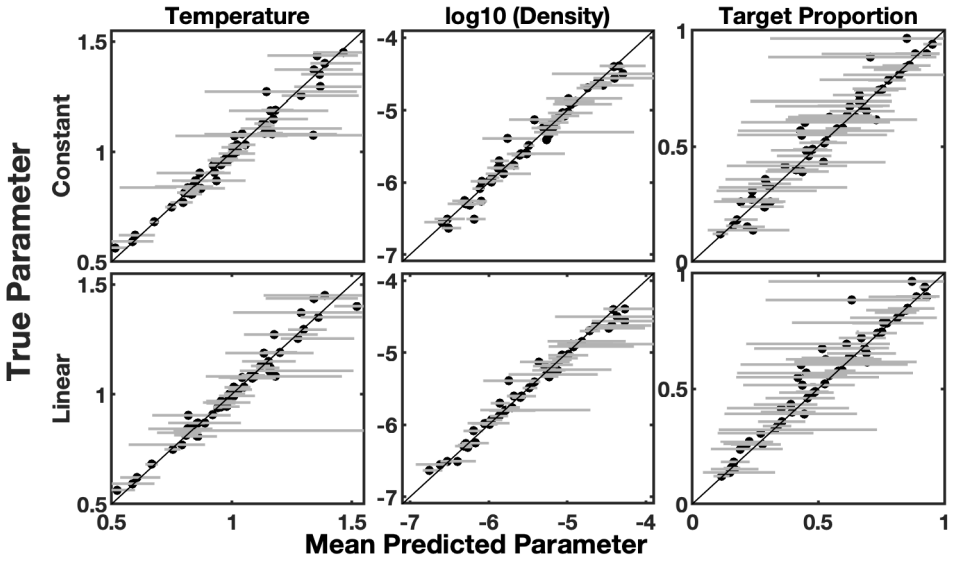

FIG. 5. *Plots of the posterior mean vs. the true value of the ATOMIC model parameters for* 12 *perfect model experiments. Each panel has* 60 *points*: 12 *experiments times five compounds. The top row represents the constant scaling factor case, the bottom row the linear scaling factor. The columns from left to right represent the parameters temperature* T, *density* ρ, *and target proportion* p *for all five compounds. The* 95% *credible regions, shown with horizontal lines, largely cover the true value of the parameter* (*when the credible interval crosses the diagonal identity line*).

and the relevant scaling factor (constant or linear) to its modeled spectrum $\eta(\theta^{\ddagger})$ to create a synthetic measured spectrum $y^{\ddagger}$. We apply our calibration approach to $y^{\ddagger}$ to verify that we can recover its original parameters $\theta^{\ddagger}$ and true generating discrepancy. We repeated this process for 12 different runs selected from interior design points that are closest to the center of the design while not having a value in the top or bottom 5% of the range for any parameter.

The results, shown in Figures 5 and 6, are encouraging. Figure 5 shows that we are accurately able to recover the true ATOMIC parameters, particularly once posterior uncertainty is taken into account. The estimates for temperature T and density ρ are particularly close to the true values, while we have wider uncertainties for our estimates of target proportion p. Figure 6 shows that we also accurately recover the discrepancy parameters used to generate the data for each of the perfect model experiments. We note that the slope parameters of the VIO spectrometer exhibit wide uncertainty (noted 44.5% in the figure). This is likely related to the paucity of data for this spectrometer; there are just 16 peaks in the VIO spectrometer, as seen in Figure 2.

4.3. *ChemCam LIBS data.* These perfect model results and our earlier diagnostics strongly suggest that we can accurately model measured ChemCam data, which we now consider. Figure 7 summarizes the posterior distributions for the ATOMIC model parameters for both the constant (black) and linear (dark gray) scaling factors. There are some visible differences between the two models. For instance, the model with the constant scaling factor favors a smaller value for the NaCl temperature parameter, while the linear scaling model is bimodal and generally flatter. There is a slight tendency for this behavior in other parameters too (e.g., the Zn parameters). This might indicate that the more flexible linear scaling model is inducing fewer constraints on the physics model, but this effect is not pronounced or always consistent (e.g., SiO_2 temperature). A few parameters (e.g., the $CaCl_2$ density and target proportion) show evidence of bimodality, indicating that ATOMIC may match experimental data with more than one parameter setting.

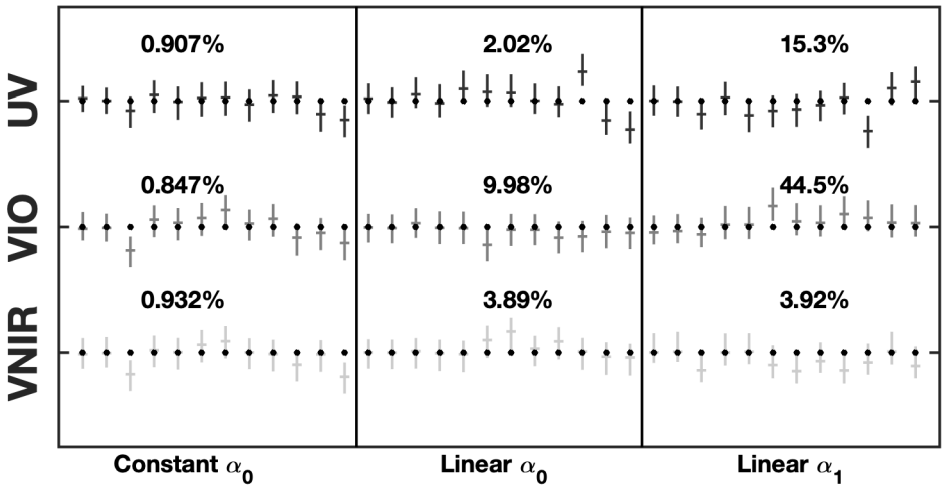

FIG. 6. *For each spectrometer and scaling factor parameter, we show the posterior mean and* 95% *credible interval for all* 12 *perfect model experiments. These intervals are calculated by*: (1) *computing the mean and* 0.025 *and* 0.975 *quantiles from the MCMC sample*, (2) *recentering the* 95% *interval by subtracting the true parameter value and* (3) *standardizing the width of the interval by the average of the* 12 *standard deviations from each MCMC sample. For most experiments, the true scaling parameters are accurately recovered, as indicated by the intervals covering the black dots. To quantify the size of our uncertainty, we report the average length of the* 95% *interval over the* 12 *experiments as a percentage of the true parameter value. The slope parameter for the VIO spectrometer has the largest uncertainty*; *the average length of its* 95% *interval is* 44.5% *of the true value. This uncertainty could be due to the relative paucity of data for this spectrometer, as shown in Figure* 2.

The difference between the two scaling-factor models is more pronounced for the discrepancy parameters shown Figure 8. For the UV and VNIR spectrometers the constant and linear scaling factor models are similar. The posterior distributions for the intercept term in the linear model (bottom-left panel) are close to the posterior distributions for the respective constant terms (top-left panel). The slope parameters for those two spectrometers (right panel) both span the zero line. The VIO spectrometer, however, seems to strongly favor a linear scaling factor. The intercept term (bottom-left panel) is noticeably higher than the constant term (top-left panel). Further, the slope parameter appears to be significantly negative with no mass near zero. This is a surprise, given the scientists' expectation of a constant shift.

Figure 9 shows calibrated predictions from the emulator (i.e., we use the draws of the parameters from the posterior distribution as inputs to the emulator and add draws from the posterior of the scaling discrepancy) for both scaling models for a subset of peaks from KCl (top row), SiO_2 (second row) and $CaCl_2$ (bottom two rows). The gray bands show the 95% credible intervals for the posterior predictions of the data used in calibration. The black lines show the experimental data. We are looking to see the colored bands mostly cover the black lines. The left panels show the predictions using the constant scaling factor, and the right panels show the predictions using the linear scaling factor.

Overall, the calibrated model does well in capturing the data. In general, the calibrated predictions match the overall level of the response. The calibrated predictions sometimes miss at the ends of the peaks (e.g., O-795.08). This might indicate that ATOMIC has trouble replicating the shoulders of the peaks. This suggests that further discrepancy modeling may be helpful to correct this behavior.

The predictions for most peaks are similar between the two models, but the calcium peaks in the third row demonstrate the significance of the linear scaling factor. The predictions in blue for the linear discrepancy are closer to both calcium peaks than the corresponding predictions for the constant discrepancy. This suggests that the linear discrepancy model for this

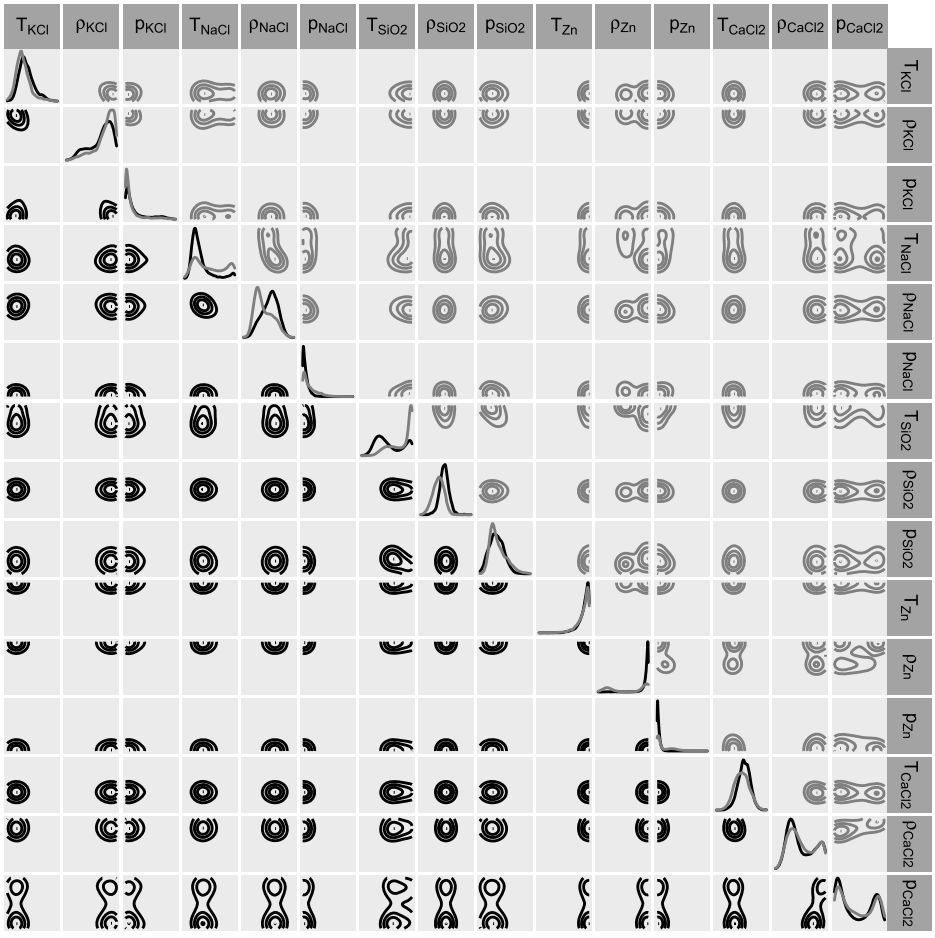

FIG. 7. *Posterior distributions for the ATOMIC model parameters for the constant* (*black*) *and linear* (*dark gray*) *scaling factors. The parameters are temperature* T, *density* ρ *and target proportion* p *for each of the five compounds*: *KCl*, *NaCl*, SiO_2, *Zn and* $CaCl_2$.

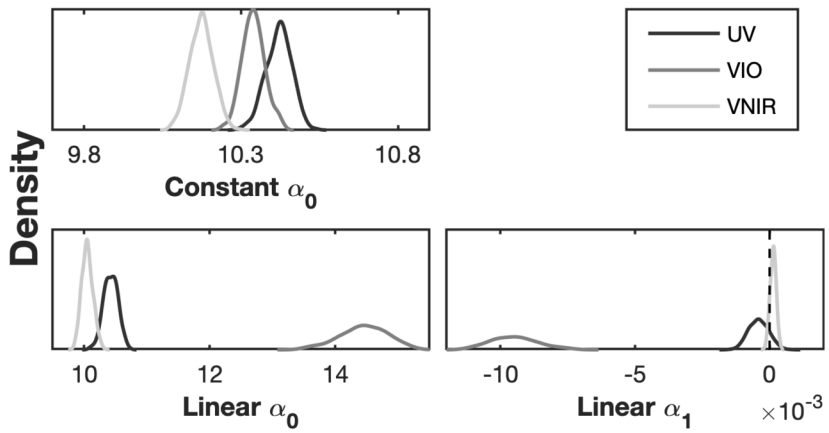

FIG. 8. *Posterior distributions for the constant scaling parameter* (*top row*) *and for the intercept term* (*bottom left*) *and slope term* (*bottom right*) *in the linear scaling model. The UV and VNIR constant scaling parameters are similar in value to the intercepts for the linear scaling model, and the slope parameters span the zero line. In contrast, the VIO spectrometer shows a strong linear effect. The intercept term for VIO is higher than the constant scaling parameter, and the posterior for the slope is significantly negative.*

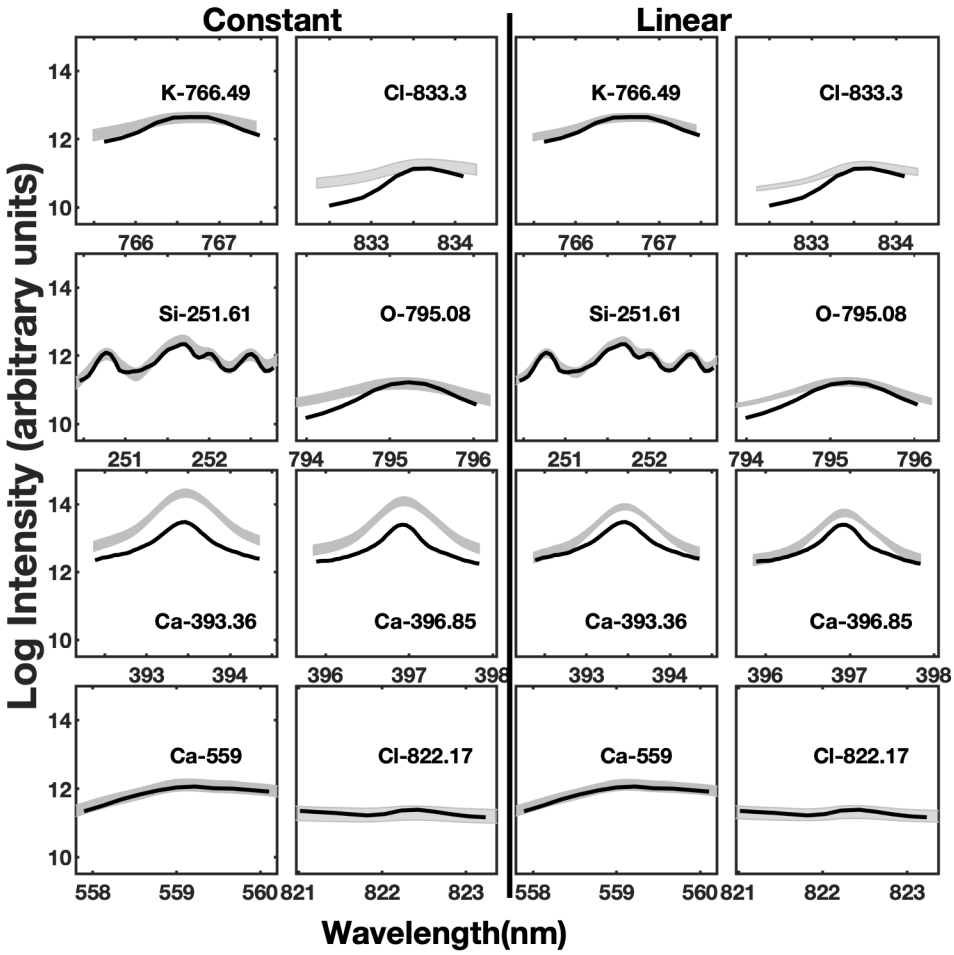

FIG. 9. *Plots of the posterior calibration model predictions and* 95% *credible intervals* (*in shaded gray*) *for the constant scaling factor approach* (*left*) *and linear scaling factor approach* (*right*). *Black lines in each panel show the measured spectral peaks. While we analyzed all the selected peaks for all five compounds, in this figure we highlight just a few from KCl* (*top row*), SiO_2 (*second row*) *and* $CaCl_2$ (*bottom two rows*). *The panel text indicate the element and center wavelength of the peak.*

spectrometer does a better job of simultaneously capturing these peaks at low wavelengths and the rest of the calcium peaks at higher wavelengths in this spectrometer (see Figure 2 for the locations of the remaining peaks). The location of the constant discrepancy seems to be dominated by these other peaks, while the linear discrepancy can capture a trend over the data. Both models have difficulty in capturing these peaks, but the linear scaling model does a much better job. Calcium has 12 of the 16 selected peaks in the VIO spectrometer which is the only instrument to show a significant linear response. Because calcium peaks dominate this spectrometer, this effect may have more to do with this compound than the instrument. As shown in Figure 10, the simulated spectra for $CaCl_2$ have broad peaks that decay slowly over the range of this spectrometer. The observed spectrum does not show this behavior, and the linear scaling factor helps to correct the difference. There seems to be a real effect here worth investigating further, whether it's due to the spectrometer's characteristics or to ATOMIC's ability to simulate this compound.

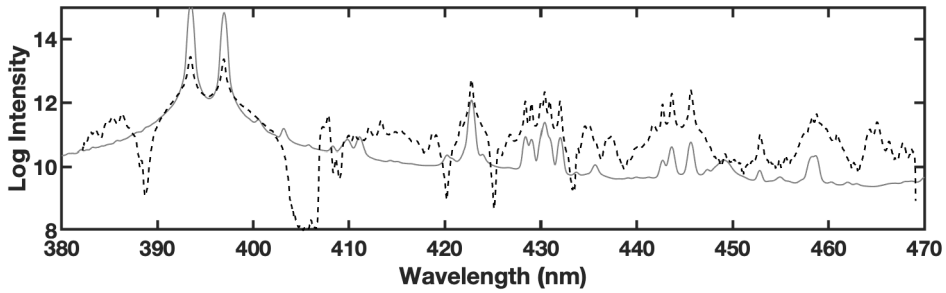

FIG. 10. *Comparison of experimental data* (*dotted black*) *and a simulation* (*solid gray*) *for* $CaCl_2$ *on the VIO spectrometer. Here, we shifted the simulation using the mean intercept, but not the linear term, for the VIO linear scaling model. The ATOMIC simulation shows a slow decay from the prominent calcium peaks between* 390 *and* 400 *nm. This decay is not evident in the experimental data. This difference may be the major contributor to the significant linear effect we found for this spectrometer. Since calcium dominates this spectrometer, the linear scaling effect may be the result of ATOMIC's modeling of* $CaCl_2$ *and not intrinsic to the spectrometer itself.*

5. Discussion. Because of the potential for confounding with the physics model (Brynjarsdóttir and O'Hagan (2014)), building an effective discrepancy model is a difficult process. Careful construction of the modeling class and prior distribution is vital to getting valid solutions for calibration. This is especially true in the case when the scaling factor discrepancy doesn't necessarily represent missing physics. In our case, while we have a few suggestions that ATOMIC *might* not be correctly modeling the underlying physics of LIBS in a few cases, such as the concerns we described with CaCl2 and with the shoulders of the oxygen peaks, what we are really missing is a forward model of the detector. Treating this as a strongly constrained discrepancy lets us solve the calibration problem without the need for this detector model.

As discussed, we have entertained two fairly simple scaling-factor models. One could consider a more general function of wavelength, $f(\omega)$. For instance, this could be crafted to account for edge effects within each spectrometer. Other physics needs can also be represented here. As before, the main effects of the physics parameters should be examined to avoid confounding. The analysis can also be considered on the original scale as opposed to the log scale that we use here. In this case the model becomes $y^{\text{obs}} = e^{f(\omega)} \odot \eta(\theta) + \epsilon$,where $\odot$ indicates pointwise multiplication. As discussed earlier, this is similar to the initial model development described in Kennedy and O'Hagan (2001). Our case is somewhat easier since our multivariate output makes the problem more identifiable. We could also consider a hierarchical approach to either connect discrepancies across spectrometers or to allow scaling to vary somewhat across compounds.

In the future, we will consider compounds made of several elements instead of the simple two-element compounds considered here. This work will require the estimation of a new set of parameters, the fraction of each element in the target. This aspect of estimation is a key component of addressing the science mission of ChemCam. It remains to be seen how this new parameter space will interact with potential scaling issues. The current work lays the groundwork for this larger goal.

Acknowledgments. Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20180097ER.

REFERENCES

BRYNJARSDÓTTIR, J. and O'HAGAN, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Probl.* **30** 114007, 24. MR3274591 https://doi.org/10.1088/0266-5611/30/11/114007

COLGAN, J., JUDGE, E., KILCREASE, D. and BAREFIELD, J. II (2014). Ab-initio modeling of an iron laser-induced plasma: Comparison between theoretical and experimental atomic emission spectra. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **97** 65–73.

COLGAN, J., JUDGE, E., JOHNS, H., KILCREASE, D., BAREFIELD, J. II, MCINROY, R., HAKEL, P., WIENS, R. and CLEGG, S. (2015). Theoretical modeling and analysis of the emission spectra of a ChemCam standard: Basalt BIR-1A. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **110** 20–30.

COWAN, R. D. (1981). *The Theory of Atomic Structure and Spectra*. Number 3. Univ. California Press, Berkeley.

FONTES, C., ZHANG, H., ABDALLAH, J. JR., CLARK, R., KILCREASE, D., COLGAN, J., CUNNINGHAM, R., HAKEL, P., MAGEE, N. et al. (2015). The Los Alamos suite of relativistic atomic physics codes. *J. Phys., B At. Mol. Opt. Phys.* **48** 144014.

GATTIKER, J., MYERS, K., WILLIAMS, B., HIGDON, D., CARZOLIO, M. and HOEGH, A. (2016). Gaussian process-based sensitivity analysis and Bayesian model calibration with GPMSA. In *Handbook of Uncertainty Quantification* 1–41.

GENOVESE, C. R., NOLL, D. C. and EDDY, W. F. (1997). Estimating test-retest reliability in functional MR imaging. I: Statistical methodology. *Magn. Reson. Med.* **38** 497–507. https://doi.org/10.1002/mrm.1910380319

GRAY, G. III (2000). Classic split-hopkinson pressure bar testing. In *ASM Handbook* **8** 462–476.

GU, M. and WANG, L. (2018). Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1555–1583. MR3875809 https://doi.org/10.1137/17M1159890

HIGDON, D., GATTIKER, J., WILLIAMS, B. and RIGHTLEY, M. (2008). Computer model calibration using high-dimensional output. *J. Amer. Statist. Assoc.* **103** 570–583. MR2523994 https://doi.org/10.1198/016214507000000888

JOSEPH, V. R. and MELKOTE, S. N. (2009). Statistical adjustments to engineering models. *J. Qual. Technol.* **41** 362–375.

JUDGE, E. J., COLGAN, J., CAMPBELL, K., BAREFIELD, J. E. II, JOHNS, H. M., KILCREASE, D. P. and CLEGG, S. (2016). Theoretical and experimental investigation of matrix effects observed in emission spectra of binary mixtures of sodium and copper and magnesium and copper pressed powders. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **122** 142–148.

KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398 https://doi.org/10.1111/1467-9868.00294

LEE, H., KASHYAP, V. L., VAN DYK, D. A., CONNORS, A., DRAKE, J. J., IZEM, R., MENG, X.-L., MIN, S., PARK, T. et al. (2011). Accounting for calibration uncertainties in X-ray analysis: Effective areas in spectral fitting. *Astrophys. J.* **731** 126.

MAGEE, N. H., ABDALLAH, J., COLGAN, J., HAKEL, P., KILCREASE, D. P., MAZEVET, S., SHERRILL, M., FONTES, C. J. and ZHANG, H. (2004). Los Alamos Opacities: Transition from LEDCOP to ATOMIC. *AIP Conf. Proc.* **730** 168–179.

MARMIN, S. and FILIPPONE, M. (2018). Variational calibration of computer models. Preprint. Available at arXiv:1810.12177.

MAURICE, S., WIENS, R., SACCOCCIO, M., BARRACLOUGH, B., GASNAULT, O., FORNI, O., MANGOLD, N., BARATOUX, D., BENDER, S. et al. (2012). The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Science objectives and mast unit description. *Space Sci. Rev.* **170** 95–166.

MENG, X.-L. (2018). Conducting highly principled data science: A statistician's job and joy. *Statist. Probab. Lett.* **136** 51–57. MR3806837 https://doi.org/10.1016/j.spl.2018.02.053

MICHEL, A. P. and CHAVE, A. D. (2007). Analysis of laser-induced breakdown spectroscopy spectra: The case for extreme value statistics. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **62** 1370–1378.

NASA (2019). Curiosity rover mission overview. Available at https://mars.nasa.gov/msl/mission/overview/.

PLUMLEE, M. (2017). Bayesian calibration of inexact computer models. *J. Amer. Statist. Assoc.* **112** 1274–1285. MR3735376 https://doi.org/10.1080/01621459.2016.1211016

RAY, K. G. and MCCREERY, R. L. (1997). Simplified calibration of instrument response function for Raman spectrometers based on luminescent intensity standards. *Appl. Spectrosc.* **51** 108–116.

SELLERS, K. F., MIECZNIKOWSKI, J., VISWANATHAN, S., MINDEN, J. S. and EDDY, W. F. (2007). Lights, camera, action! Systematic variation in 2-D difference gel electrophoresis images. *Electrophoresis* **28** 3324–3332.

SJUE, S., AHRENS, J., BISWAS, A., FRANCOM, D., LAWRENCE, E., LUSCHER, D. and WALTERS, D. J. (2020). Fast strength model characterization using Bayesian statistics. In *Proceedings of the* 21*st Bienniel Conference of the APS Topical Group on Shock Compression of Condensed Matter* (*SHOCK*19).

STOVER, F. S. and BRILL, R. V. (1998). Statistical quality control applied to ion chromatography calibrations. *J. Chromatogr. A* **804** 37–43.

VAN DYK, D. A., CONNORS, A., KASHYAP, V. L. and SIEMIGINOWSKA, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophys. J.* **548** 224.

WIENS, R. C., MAURICE, S., BARRACLOUGH, B., SACCOCCIO, M., BARKLEY, W. C., BELL, J. F., BENDER, S., BERNARDIN, J., BLANEY, D. et al. (2012). The ChemCam instrument suite on the Mars Science Laboratory (MSL) rover: Body unit and combined system tests. *Space Sci. Rev.* **170** 167–227.

WIENS, R., MAURICE, S., LASUE, J., FORNI, O., ANDERSON, R., CLEGG, S., BENDER, S., BLANEY, D., BARRACLOUGH, B. et al. (2013). Pre-flight calibration and initial data processing for the ChemCam laser-induced breakdown spectroscopy instrument on the Mars Science Laboratory rover. *Spectrochim. Acta, Part B: Atom. Spectrosc.* **82** 1–27.