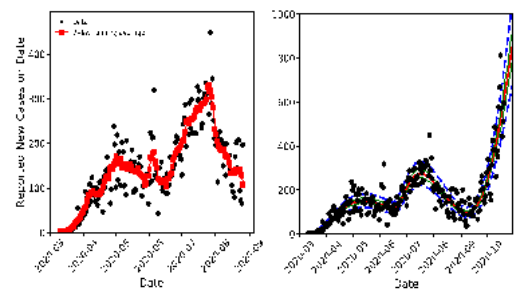
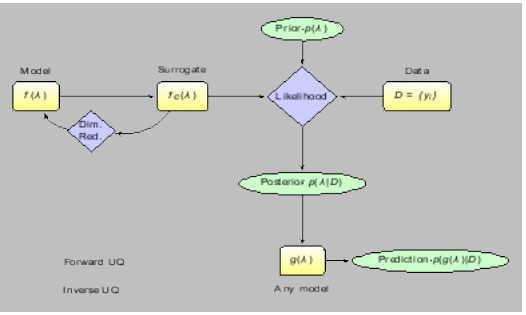
SAND2020-11793PE





Sandia National Laboratories

Exceptional service in the national interest

Parameter Inference and Forward Propagation of Uncertainties in Computational Models

Application to Covid-19 Epidemic

Cosmin Safta

(csafta@sandia.gov)

University of Michigan
October 27, 2020

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

Sandia National Laboratories

- Probabilistic Framework for Model Calibration and Predictive Assessment
  - Data, Models, Bayes' Rule
- Practical Application
  - Modeling the Covid-19 Epidemic
  - Interatomic Potential Models for Binary Alloys
  - Energy Exascale Earth System (E3SM) – Land Model Component
- Brief Description of Employment Opportunities at Sandia National Labs

Acknowledgements

Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Science,Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC).

Contract

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. The views expressed in this talk do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Disclaimer

The views expressed in this talk do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Motivation: Enable Predictive Simulations

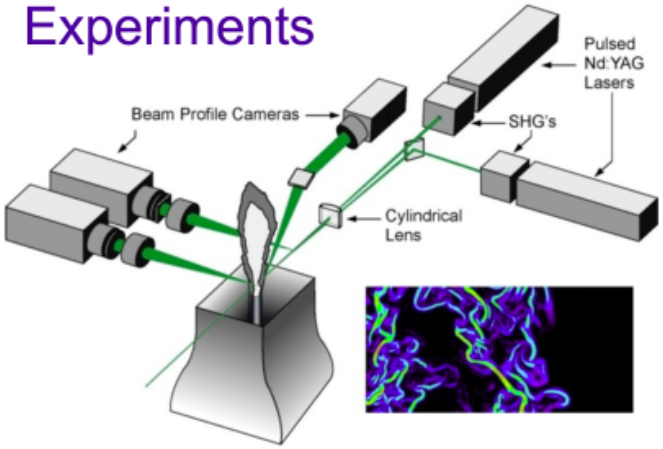
Sandia National Laboratories

Theory (incomplete)

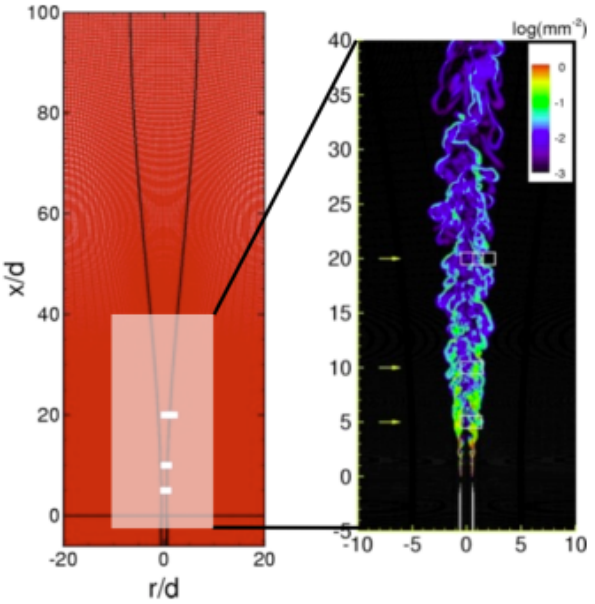
$$F(u, \lambda) = 0$$

+

Experiments



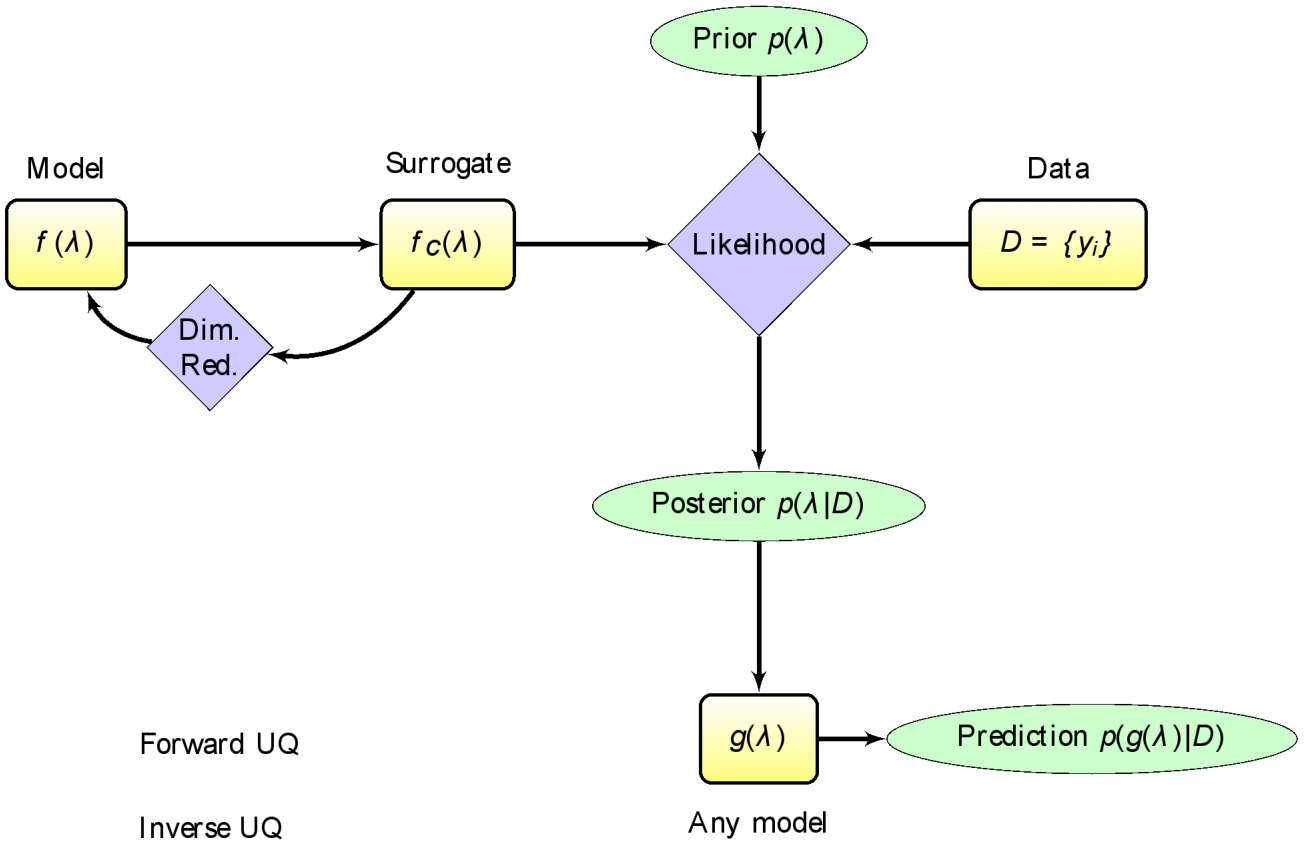
Predictive Simulation



Uncertainty Sources

- Model parameters
- Initial/boundary conditions
- Intrinsic stochasticity
- Model geometry/structure
- Data noise
- Numerical errors, too

An Example Workflow



Combine Models and Experiments in a Statistical Framework

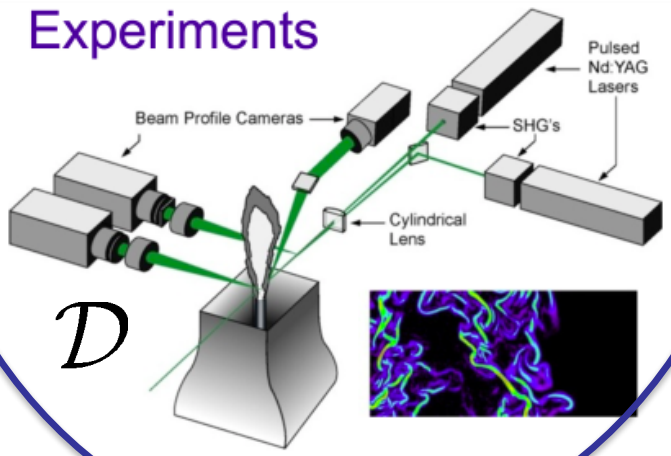
Sandia National Laboratories

Models (incomplete)

$$F(u, \lambda) = 0$$

+

Experiments



Bayes' rule:

$$\overbrace{P(\lambda|\mathcal{D})}^{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\lambda)}^{\text{Likelihood}}\overbrace{P(\lambda)}^{\text{Prior}}}{\underbrace{P(\mathcal{D})}_{\text{Evidence}}} \propto P(\mathcal{D}|\lambda)P(\lambda)$$

- Update prior distribution/knowledge about parameter λ to posterior distribution given data $\mathcal{D}$, using likelihood function

$$\mathcal{L}_{\mathcal{D}}(\lambda) = P(\lambda|\mathcal{D})$$

- ***Data*** – measurements of some quantities of interest
- ***Evidence*** – can be seen as a normalizing term

The prior distribution represents prior information about the inferred quantities

Sandia National Laboratories

$$\overbrace{P(\lambda|\mathcal{D})}^{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\lambda)}^{\text{Likelihood}}\overbrace{P(\lambda)}^{\text{Prior}}}{\underbrace{P(\mathcal{D})}_{\text{Evidence}}} \propto P(\mathcal{D}|\lambda)P(\lambda)$$

- Based on prior data, literature, or expert opinion
- Prior distribution helps to keep inference well defined, e.g. if quantity needs to remain positive
- If not much data available, posterior will be strongly influenced by the prior
- When a lot of data available (and it is relevant to the model at hand), data will have predominant influence on posterior
- Prior is both powerful and dangerous
- If no prior information is available, non-informative priors can be used
  - e.g. uniform over some physical range

The likelihood function measures goodness-of-fit

$$\overbrace{P(\lambda|\mathcal{D})}^{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\lambda)}^{\text{Likelihood}}\overbrace{P(\lambda)}^{\text{Prior}}}{\underbrace{P(\mathcal{D})}_{\text{Evidence}}} \propto P(\mathcal{D}|\lambda)P(\lambda)$$

- The key component that connects the model inputs to measured QoIs
- Statistical model to account for disagreement between model and data
  - Common case is i.i.d. Gaussian measurement noise in each data point

$$\mathcal{L}_{\mathcal{D}}(\lambda) = P(\mathcal{D}|\lambda) = \frac{1}{(2\pi)^{N/2}\sigma^N}\exp\left(-\sum_{i=1}^{N}\frac{(d_i - f_i(\lambda))^2}{2\sigma^2}\right)$$

- If the model itself is uncertain, then the noise model needs to reflect that

The posterior contains updated knowledge about inferred parameters

$$\overbrace{P(\lambda|\mathcal{D})}^{\text{Posterior}} = \frac{\overbrace{P(\mathcal{D}|\lambda)}^{\text{Likelihood}}\overbrace{P(\lambda)}^{\text{Prior}}}{\underbrace{P(\mathcal{D})}_{\text{Evidence}}} \propto P(\mathcal{D}|\lambda)P(\lambda)$$

- Gives the inferred values of the parameters as well as their uncertainty based on all sources of uncertainty
  - The maximum value is referred to as the *Maximum A Posterior* (MAP) value
- Posterior distribution generally not analytically tractable
- Commonly people resort to sampling approaches, e.g. Markov Chain Monte Carlo to draw samples from this distribution
  - Can be used to understand correlations between model components
  - Can be fed into other models to augment model predictions with information extracted from data

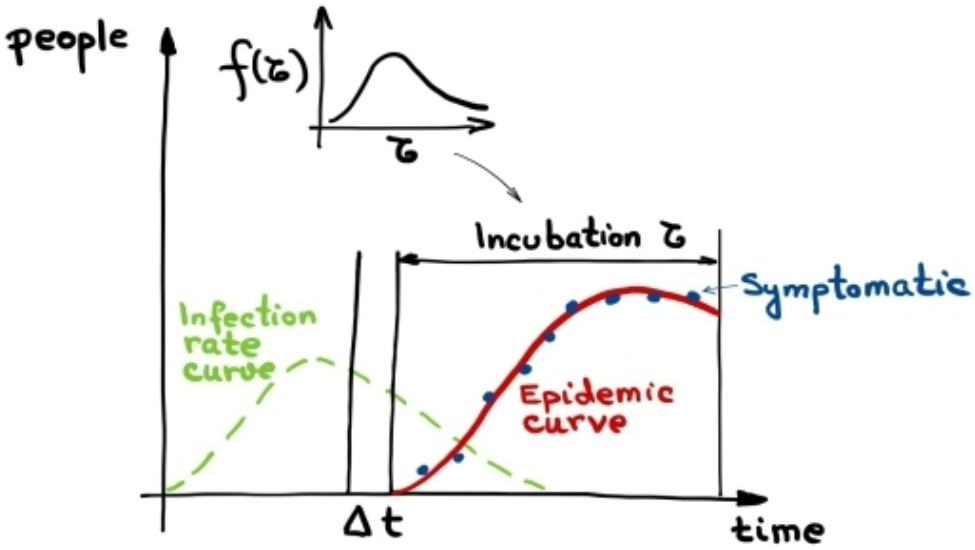
Outline

- Probabilistic Framework for Model Calibration and Predictive Assessment
  - Data, Models, Bayes' Rule
- Practical Application
  - Modeling the Covid-19 Epidemic
  - Interatomic Potential Models for Binary Alloys
  - Energy Exascale Earth System (E3SM) – Land Model Component
- Brief Description of Employment Opportunities at Sandia National Labs

Characterization of Partially Observed Epidemics: Application to Covid-19

Sandia National Laboratories

Infections map to epidemic via incubation



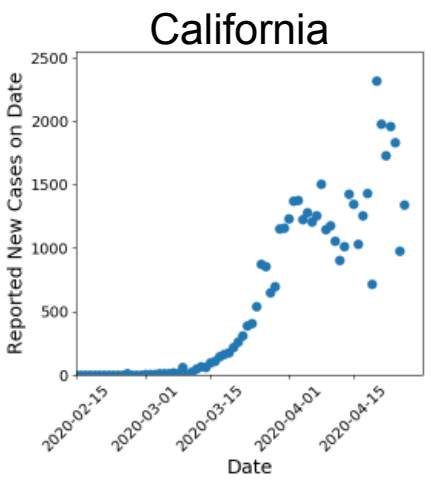
Model:

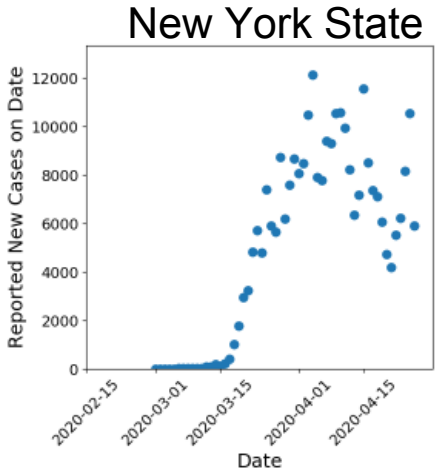
- Symptomatic cases observed on a certain day are a consequence of people infected at various times coming out of incubation and presenting symptoms
- The incubation period is drawn from COVID-19 incubation period distribution

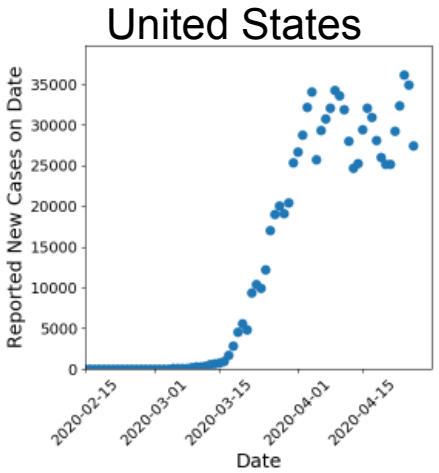
Characterization is the estimation of infection spread parameters using daily counts of symptomatic patients. The method is designed to help guide medical resource allocation in the early epoch of the outbreak.

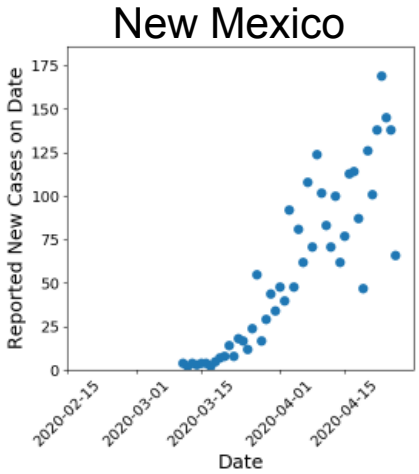
Covid-19 Data

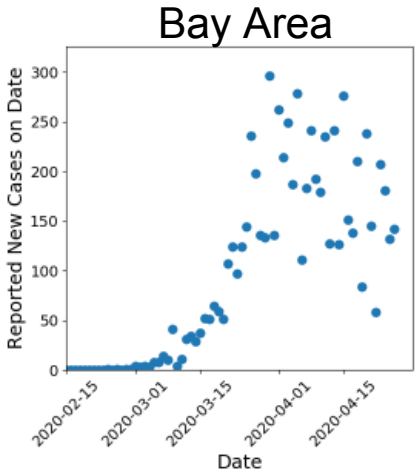
Sandia National Laboratories

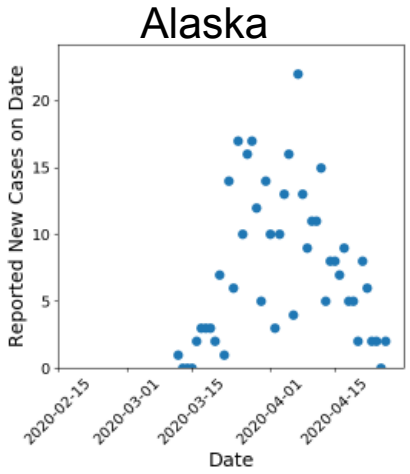












Data from NY Times and Johns Hopkins University github repositories

Model Details-1

- ***Infection Rate*** curve modeled as a Gamma distribution with unknown shape (k) and scale (θ) parameters

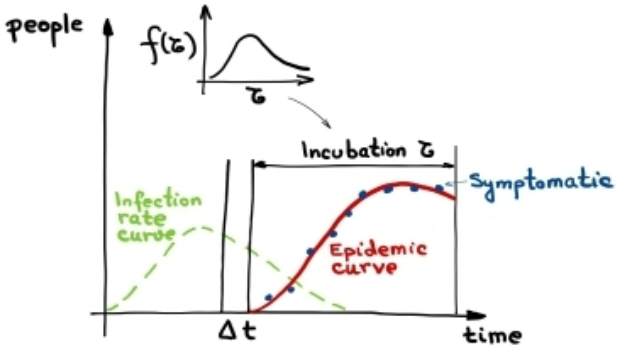
$$\mathrm{InfR}(t - t_0) \sim \Gamma(k, 1/\theta)$$

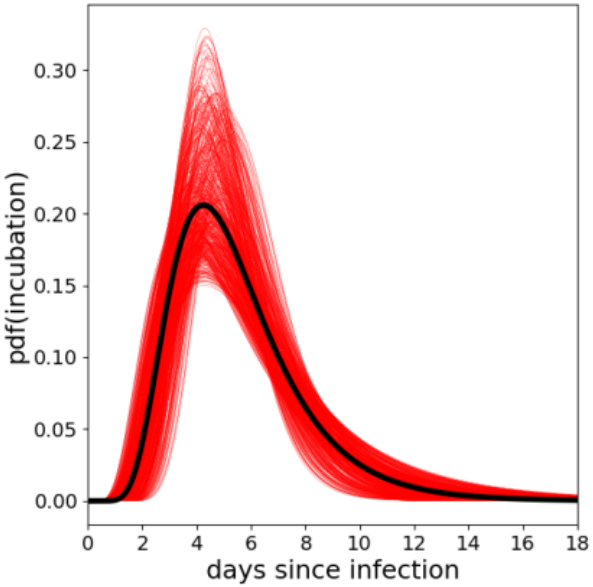
- ***Incubation Rate*** is modeled using a log-normal distribution with parameters based on published results

$$\mathrm{IncR} \sim \mathrm{Lognormal}(\mu(\xi_1), \sigma(\xi_2)^2)$$

$$\mu = 1.504 \ldots 1.755$$
$$\sigma = 0.271 \ldots 0.542$$

Lauer et al, *"The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application*" Annals of Internal Medicine, 2020





Model Details-2

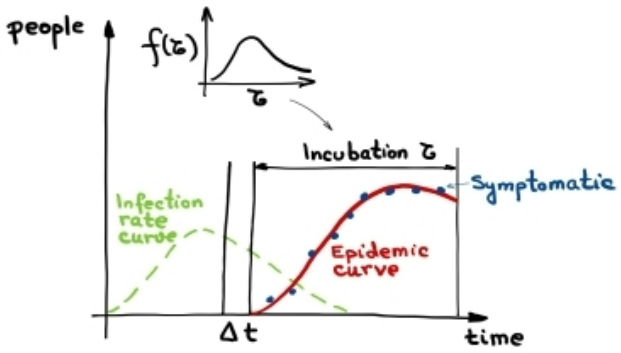
- People showing symptoms on day t_i - convolution between infection rate and incubation rate models

$$n_i(t_0, N, k, \theta, \xi_1, \xi_2) = N \int_{t_0}^{t_i} f_\Gamma(\tau - t_0; k, 1/\theta) f_{LN}(t_i - \tau; \mu(\xi_1), \sigma(\xi_2)) d\,\tau$$

- Back to Bayes' theorem

$$p(\Theta|D) \propto L_D(\Theta) \times p(\Theta)$$

$$\Theta = \{t_0, N, k, \theta, \underbrace{\ldots}_{\epsilon}\}$$



- How do we model the discrepancy between the model and the data?

Approximations for the Discrepancy Between Data and Models

Sandia National Laboratories

- Likelihood for *fixed incubation* model
  - discrepancy approximated as independent daily Gaussian discrepancies

$$L_D = \prod_{i=1}^{N_d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_{obs,i} - n_i)^2}{2\sigma_i^2}\right)$$

- Likelihood for *uncertain incubation* model

$$L_D = \prod_{i=1}^{N_d} p_i(y_{obs,i}|\Theta)$$

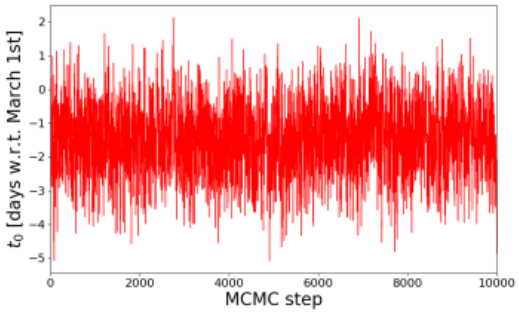
- Looked at both additive and additive/multiplicative error models

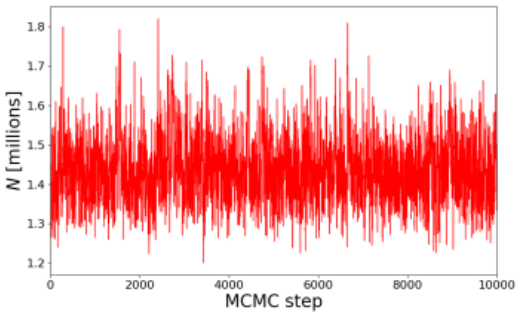
$$\sigma_i = \sigma_a + \sigma_m \times n_i$$

Negative binomial distributions, typically used in epidemiology did not yield satisfactory results for regions with a large number of cases

Sampling the Joint Posterior Distribution

- Sampling via Markov Chain Monte Carlo (MCMC)
  - Model evaluations are cheap, about 5 min for 1 million samples on my laptop



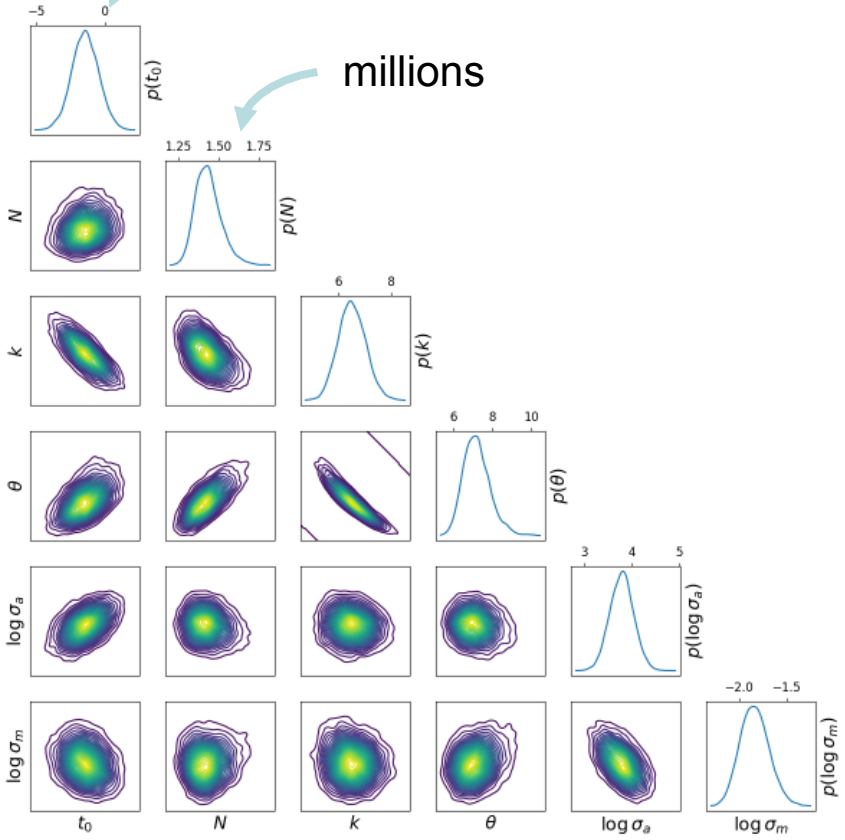


.....

Kernel Density Estimate

March 1st

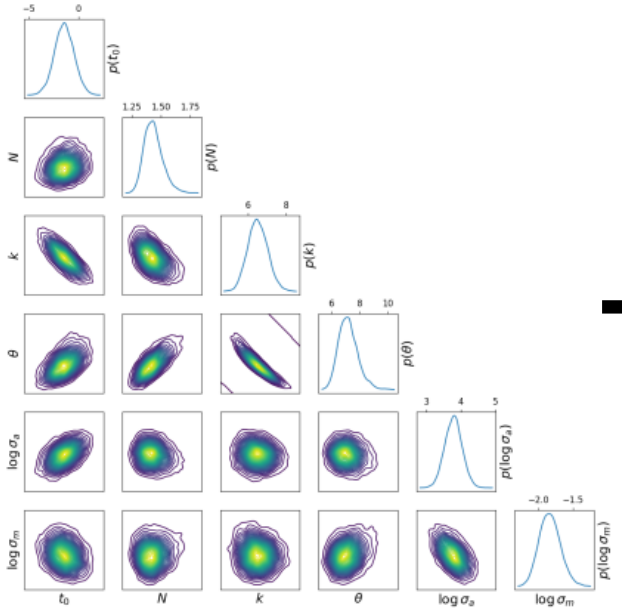
millions

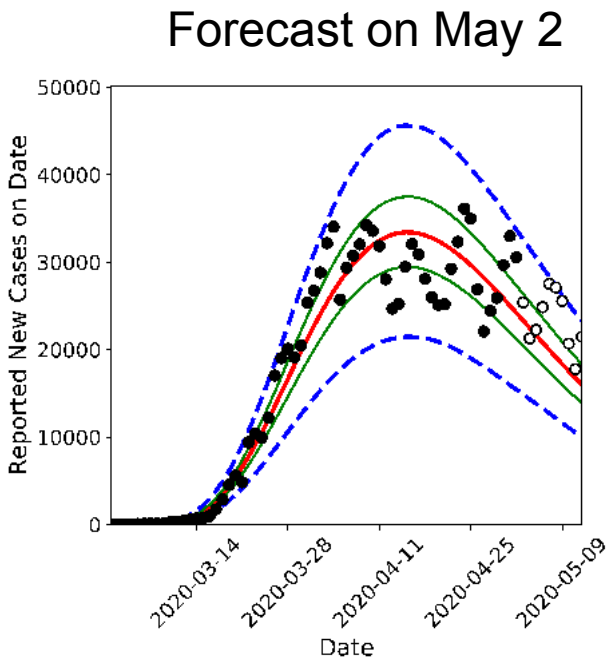


Predictive Assessment – Pushed and Posterior Predictive Distributions

- Posterior predictive distributions

$$p_{\mathrm{pp}}\left(n^{(\mathrm{pp})}|\mathcal{D}\right) = \int_{\Theta} p(n^{(\mathrm{pp})}|\Theta)\,\underbrace{p(\Theta|\mathcal{D})}_{\text{Posterior}}\,d\Theta.$$

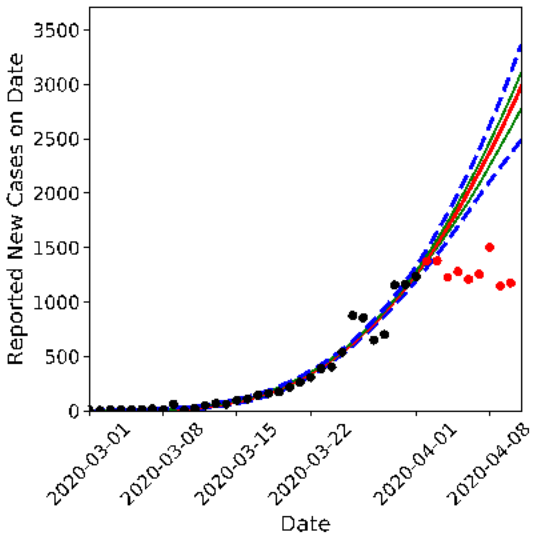




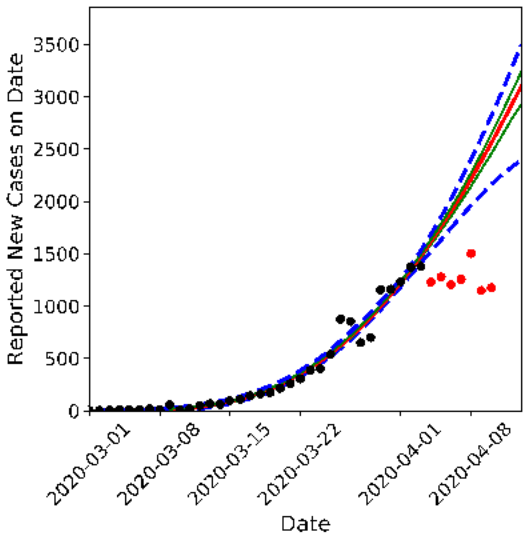
Results – Early Forecasts for CA

- This model captures stationary behavior. Any change in social dynamics will be observed approx. 5-10 days later after the median incubation period.
- The series of forecasts below for California show the impact of stay at home order issued on March 19 on *"flattening the curve"*

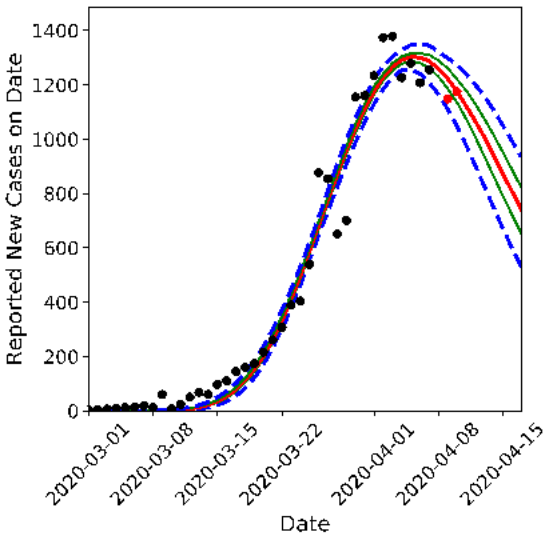
Forecast on April 1



Forecast on April 3



Forecast on April 7

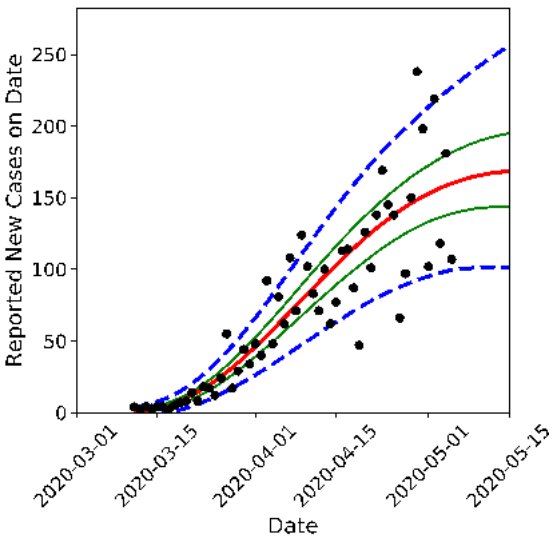


- Black symbols show data used for model inference and to generate forecasts
- Red symbols display data observed after the forecast was produced

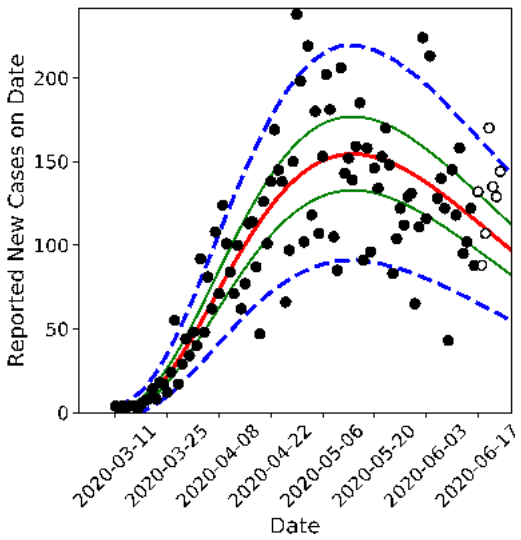
Results – New Mexico

- Results used by the NM Department of Health to assess weekly trends

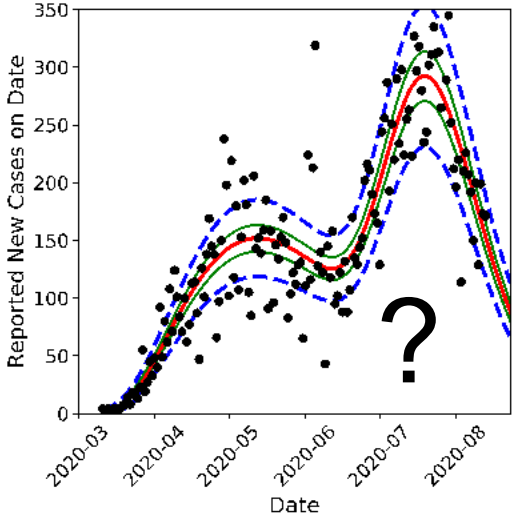
Forecast on May 5



Forecast on June 23



Forecast on August 13



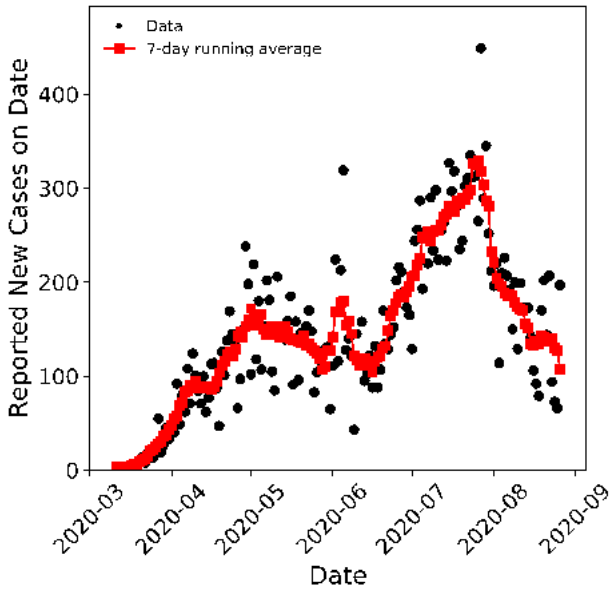
- Black symbols show data used for model inference and to generate forecasts
- White circles display data observed after the forecast was produced

Re-evaluate the Modeling Approach

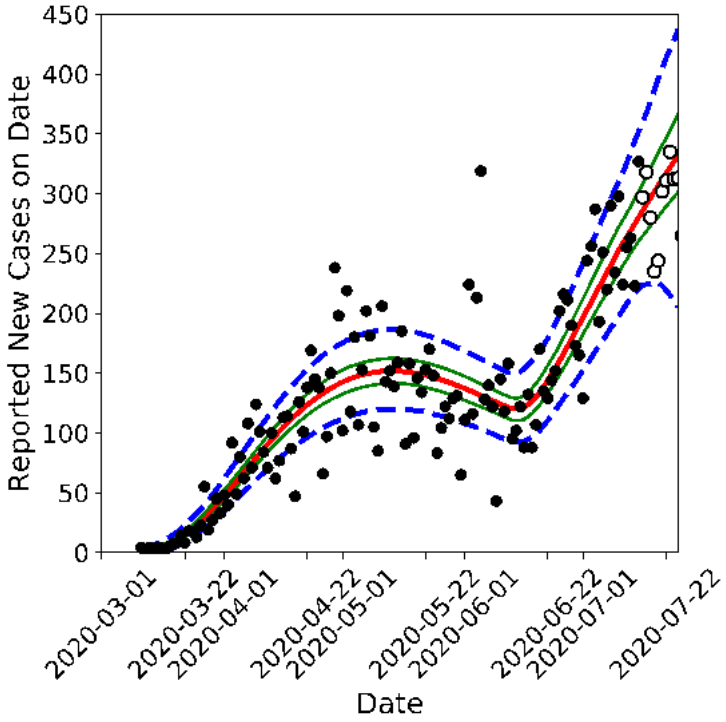
- Lets try multiple infection curves

$$n_i = \int_{t_0}^{t_i} \left(\sum_{j=1}^{K} N_j f_\Gamma(\tau - t_0 - \Delta t_j; k_j, \theta_j) \right) f_{LN}(t_i - \tau; \mu, \sigma) d\tau$$

- Data



NM Forecast on July 15 (2 waves)

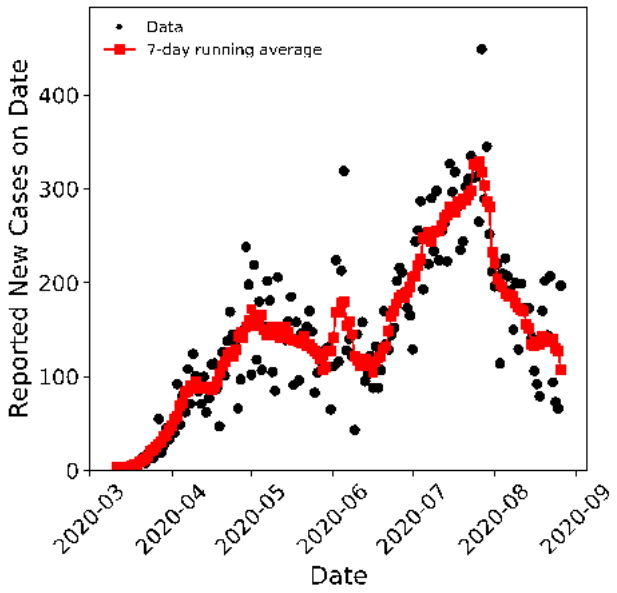


Re-evaluate the Modeling Approach

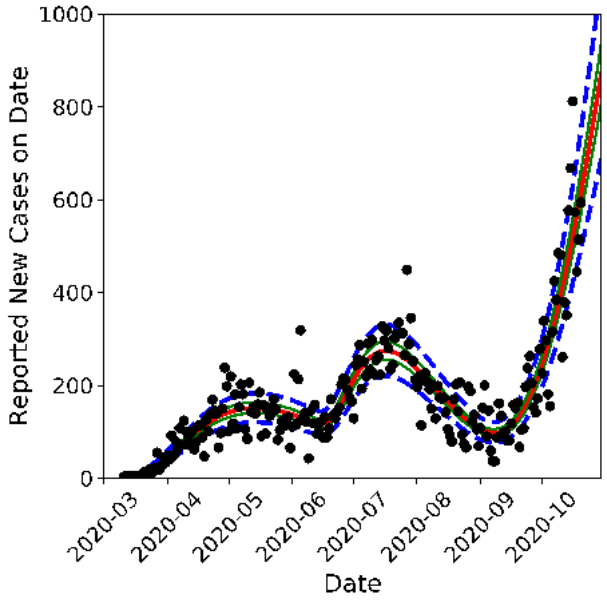
- Lets try multiple infection curves

$$n_i = \int_{t_0}^{t_i} \left(\sum_{j=1}^{K} N_j f_\Gamma(\tau - t_0 - \Delta t_j; k_j, \theta_j) \right) f_{LN}(t_i - \tau; \mu, \sigma) d\tau$$

- Data



Forecast on October 20 (3 waves)



decide on the appropriate model complexity via information criteria (with P. Blonigan, J. Ray)

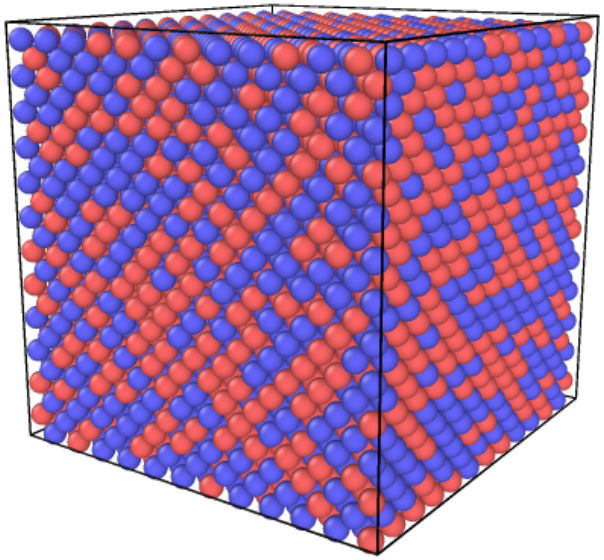
Outline

- Probabilistic Framework for Model Calibration and Predictive Assessment
  - Data, Models, Bayes' Rule
- Practical Application
  - Modeling the Covid-19 Epidemic
  - Interatomic Potential Models for Binary Alloys
  - Energy Exascale Earth System (E3SM) – Land Model Component
- Brief Description of Employment Opportunities at Sandia National Labs

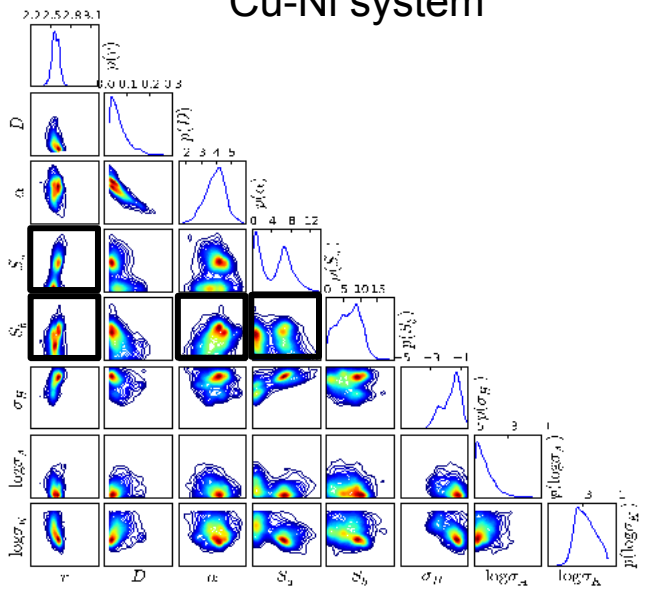
Other Applications – Modeling Interatomic Potentials for Binary Alloys

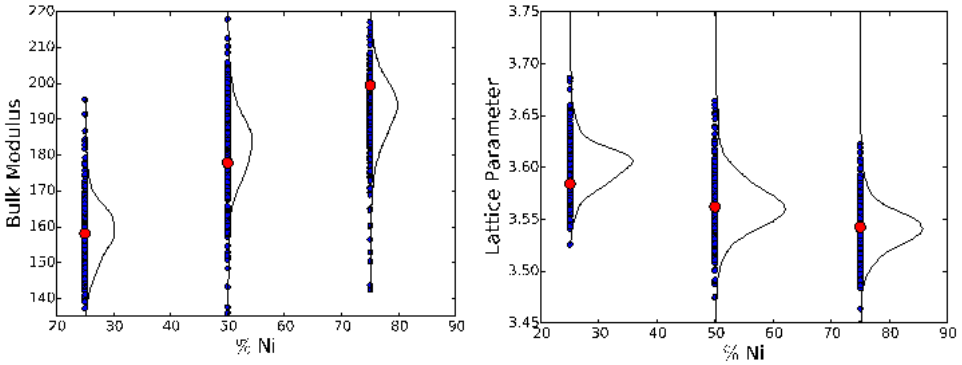
Sandia National Laboratories

- Employ molecular dynamics simulations to calibrate inter-atomic potentials for binary alloys designed for long-term nuclear waste storage



Cu-Ni system

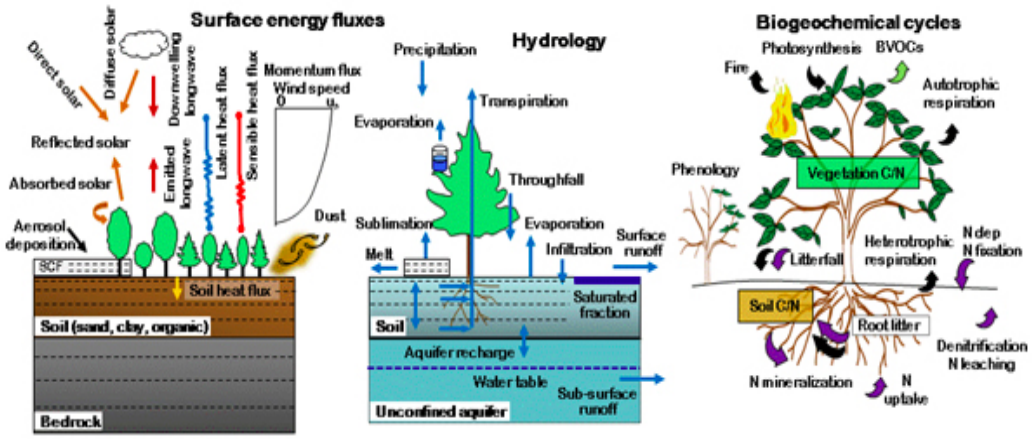


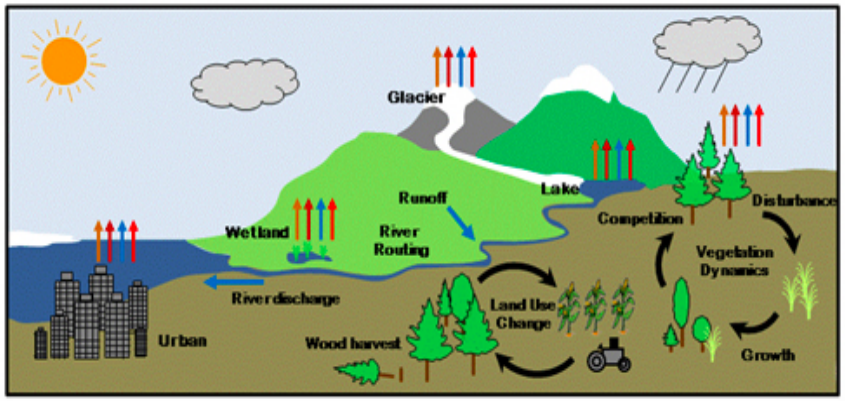


- with A. Hegde, H. Najm (SNL), W. Windl, E. Weiss (OSU)

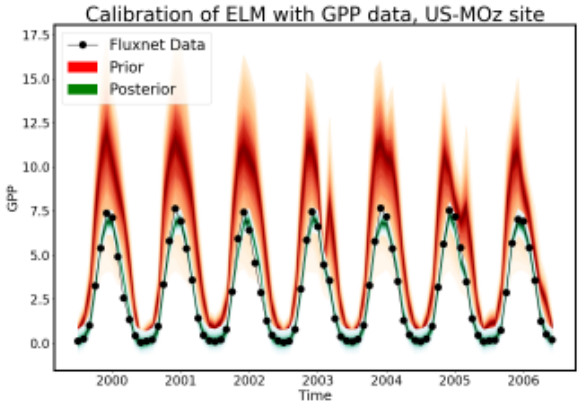
Energy-Exascale Earth System Model (E3SM) Land Component

Sandia National Laboratories





Understanding physical processes is critical to understand the climate feedbacks and their sensitivity to uncertainties in parameters and model structure



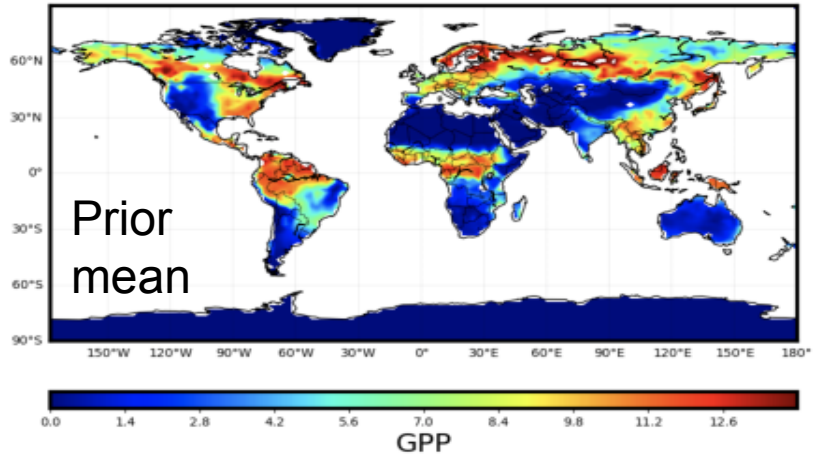
Gross Primary Production (GPP) at the AmeriFlux Missouri Ozark site (US-MOz)

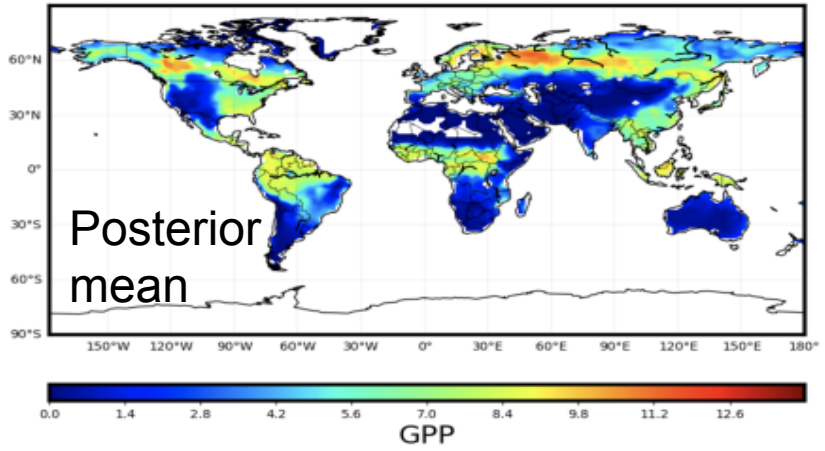
with K. Sargsyan (SNL), D. Ricciuto (ORNL)

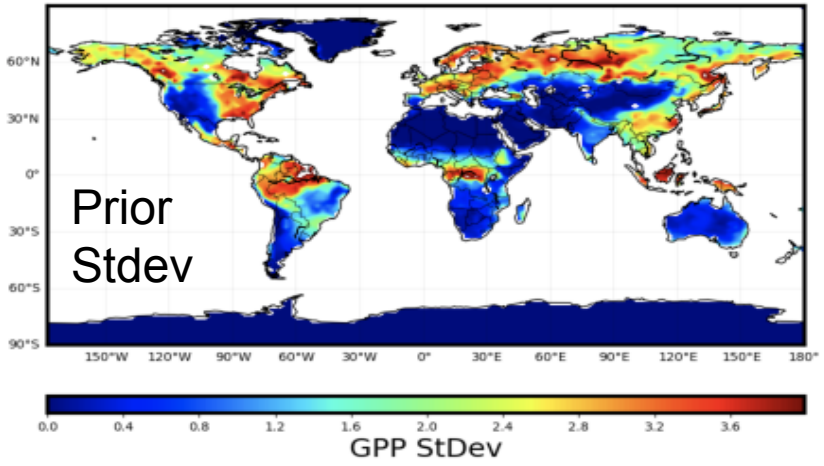
FASTMATH

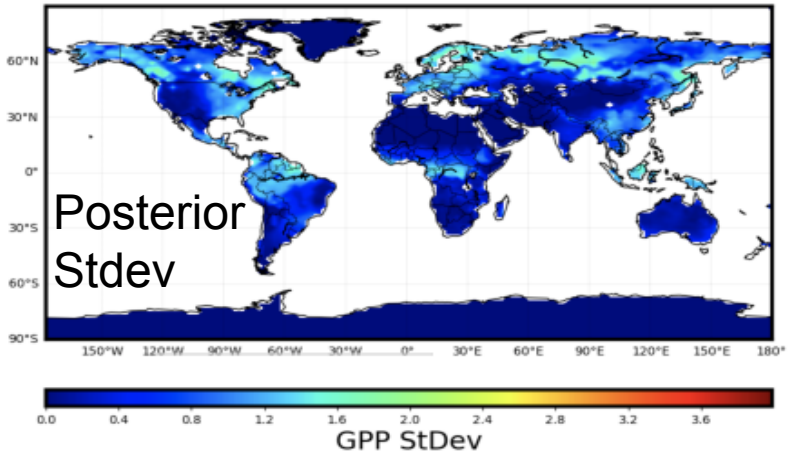
E3SM Land Model - Impact of Calibration

- Calibrated PDF using 28 US sites, used for *posterior prediction globally*
- Predictions shown for one month (July 2001)









Summary/Challenges/Work-in-progress

Sandia National Laboratories

- High-dimensionality: large number of input parameters (10s-100s) and computational expense
  - surrogate models/reduced-order models
  - hierarchy of model fidelities
- Data quality
  - modeling approaches that can operate with incomplete or corrupted data
- Our understanding of physical processes is frequently incomplete
  - design algorithms that can embed the discrepancy between models and data inside the model and the carry it in subsequent analyses

Outline

- Probabilistic Framework for Model Calibration and Predictive Assessment
  - Data, Models, Bayes' Rule
- Practical Application
  - Modeling the Covid-19 Epidemic
  - Interatomic Potential Models for Binary Alloys
  - Energy Exascale Earth System (E3SM) – Land Model Component
- Brief Description of Employment Opportunities at Sandia National Labs

Sandia Has Two Main Locations: Albuquerque (NM) & Livermore (CA)

National Nuclear Security Administration labs
Science labs
Nuclear energy lab
Environmental management lab
Fossil energy lab
Energy efficiency and renewable energy lab

Our Workforce ~14,100 employees

Sandia National Laboratories

~12,300 Regular employees

~1,800 Temporary employees, students & postdoctoral appointees

New Mexico Site:

Workforce: ~12,500

R&D employees: ~4,200

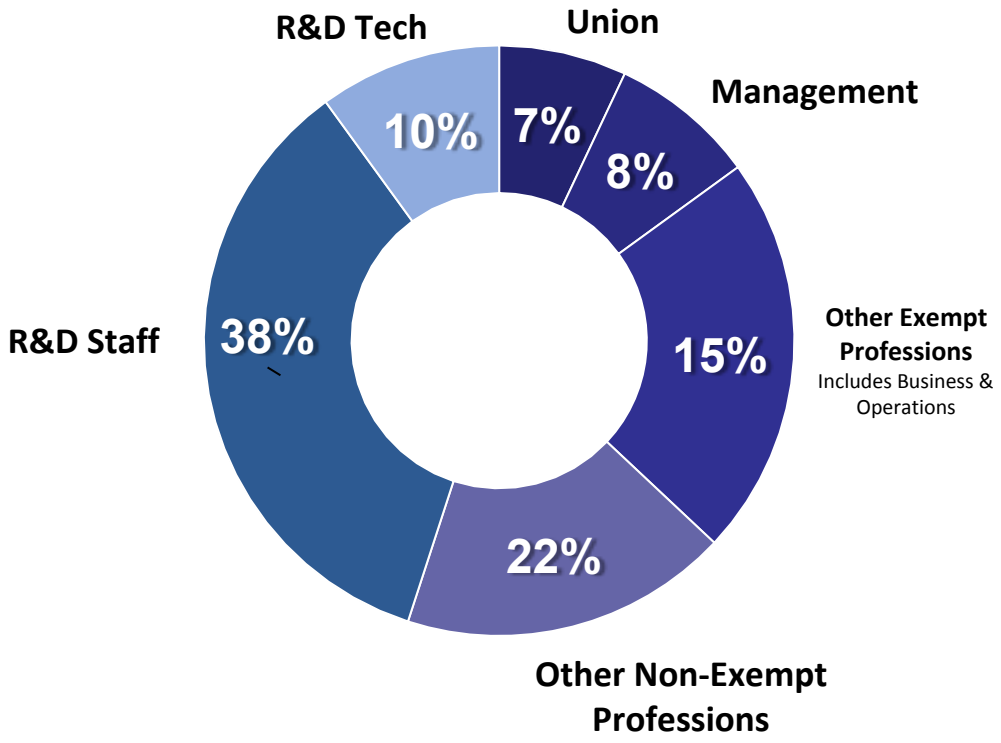
(R&D Staff & Technologists)

California Site:

Workforce : ~1,600

R&D employees: ~650

(R&D Staff & Technologists)



Internships

Encourages qualified students to develop interests in critical skills areas related to our mission, with the ultimate objective of developing our pipeline for our future. Available for Summer, Year Round and Co-op.

Eligibility Criteria

- Full-time enrollment status at an accredited school during the academic school year
- Undergraduate equivalent of 12 hours per semester
- Graduate equivalent of 9 hours per semester
- Must have a minimum cumulative GPA of 3.0 on a 4.0 scale for Technical, R&D, and Business interns; 2.5 on a 4.0 scale for Clerical and Labor interns
- Have U.S. citizenship for positions that require a security clearance or as stated in the job posting
- At least 16 years of age

http://www.sandia.gov/careers/students_postdocs/internships/index.html

Post-doc Opportunities

Key areas for post-docs at Sandia:

- Computer science/Computer Engineering
- Electrical Engineering
- Mechanical Engineering
- High-performance computing
- Microelectronics and microfluidics
- Nanotechnology
- Physics
- Chemistry/ Electro Chem
- Biosciences and biotechnology
- Radiation & electrical sciences
- Engineering sciences
- Pulsed power sciences
- Materials science & engineering

Eligibility Criteria

- A recent PhD (conferred 5 years prior to employment) or the ability to complete all PhD requirements before hire date.

http://www.sandia.gov/careers/students_postdocs/postdocs.html

Fellowship Opportunities

Sandia provides postdoctoral fellows with professional development opportunities and prepares fellows to conduct independent, groundbreaking research.

Postdoctoral Fellowships

- Harry S. Truman Fellowship
- Jill Hruby Fellowship
- John Von Neumann

http://www.sandia.gov/careers/students_postdocs/fellowships/index.html