Towards Data-Informed Climate Sciences - Leveraging Machine Learning Inferences of Satellite Observations

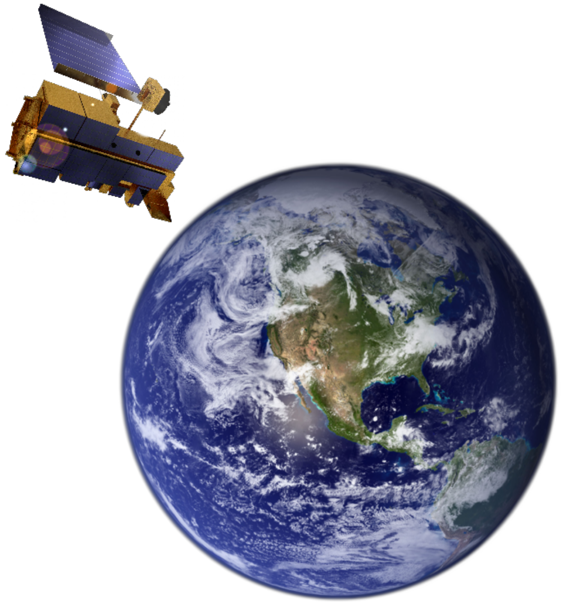
Srija Chakraborty

USRA, NASA GSFC

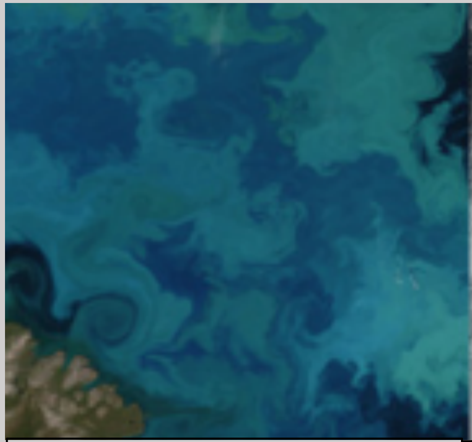
August 20, 2020

srija.chakraborty@nasa.gov
schakr34@asu.edu

Understanding Earth System Components

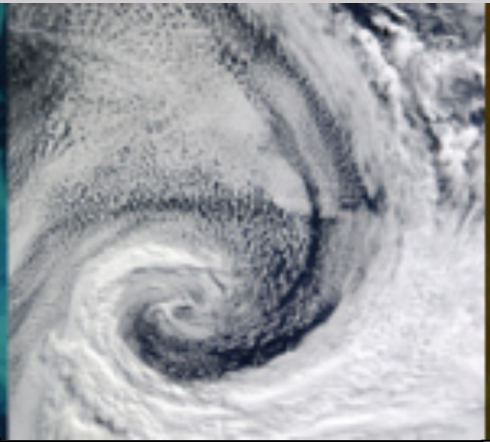


Ocean



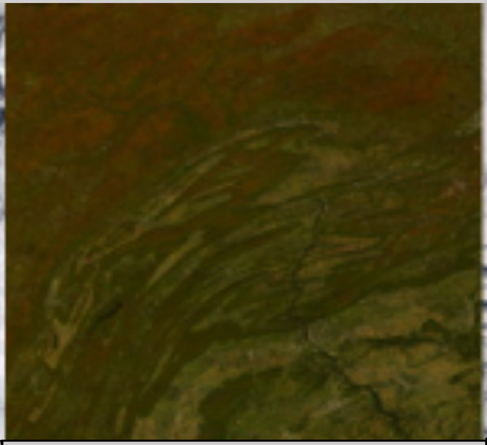
- Sea Surface Temperature
- Chlorophyll
- Particulate Organic & Inorganic Carbon
- Salinity

Atmosphere



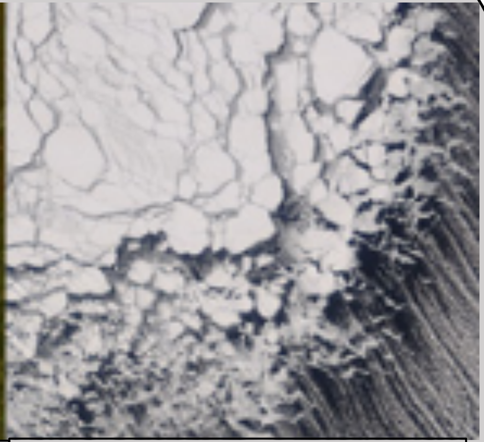
- Aerosols
- Carbon Monoxide
- Clouds
- Atmospheric Profile
- Carbon Dioxide
- Precipitable Water

Land



- Surface Reflectance & Temperature
- Land Cover
- Vegetation
- Gross Primary Production
- Thermal Anomaly
- Burned Area

Cryosphere



- Snow Cover
- Sea Ice & Ice Surface Temperature

- Varying modalities
- Varying resolution

https://terra.nasa.gov/about/terra-instruments/modis
https://modis.gsfc.nasa.gov/sci_team/

Scope:

Data Sources:

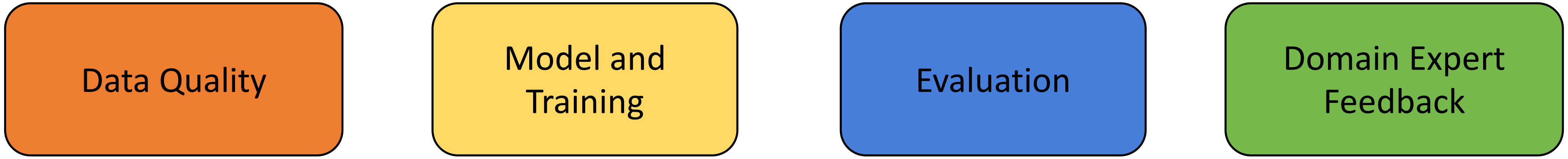
- Physics/ Climate Models
- In-Situ Observations
- Remote Sensing
- Citizen Science Records

Scope:

Data Sources:

- Physics/ Climate Models
- In-Situ Observations
- Remote Sensing
- Citizen Science Records

Challenges and Directions with Remote Sensing:



Data Quality

Data Quality from Data Generation Process

top of atmosphere observations to meaningful parameters

Top of atmosphere signal

Atmospheric gases, ozone, aerosol, trace gases, clouds

Surface

Atmospheric Correction

Input conditions, radiative transfer models, BRDF

Retrieved Surface parameters

Surface reflectance, vegetation, snow

Data quality (per-pixel) encoded

Data Quality

Data Quality from Data Generation Process

top of atmosphere observations to meaningful parameters

Top of atmosphere signal

Atmospheric gases, ozone, aerosol, trace gases, clouds

Surface

Atmospheric Correction → Retrieved Surface parameters → Data quality (per-pixel) encoded

Input conditions, radiative transfer models, BRDF

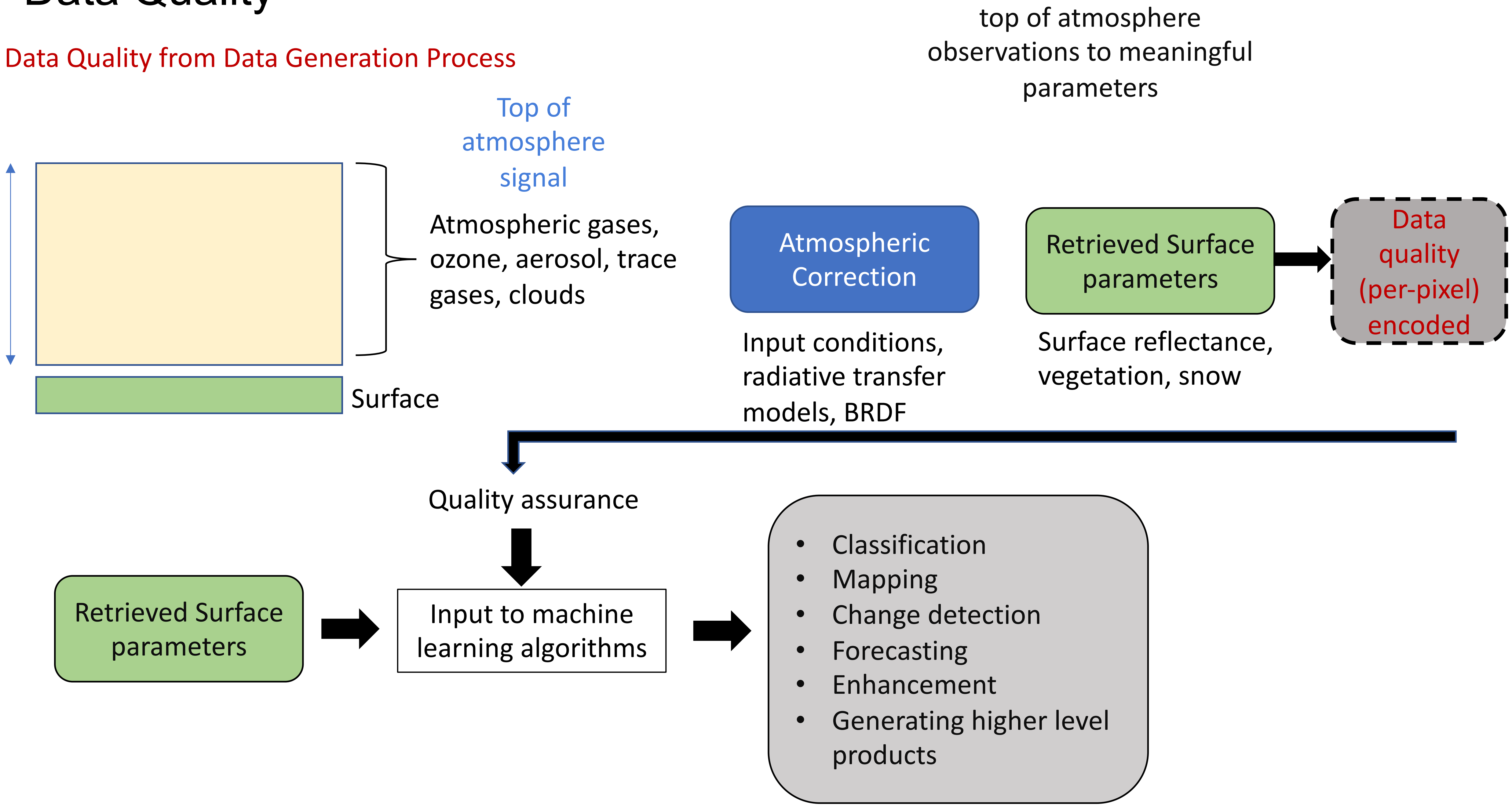
Surface reflectance, vegetation, snow

Retrieved Surface parameters → Input to machine learning algorithms →

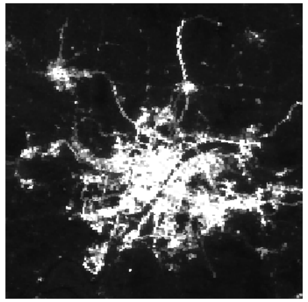
- Classification
- Mapping
- Change detection
- Forecasting
- Enhancement
- Generating higher level products

Data Quality

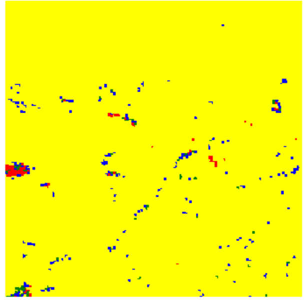
Data Quality from Data Generation Process



Impact of Data Quality



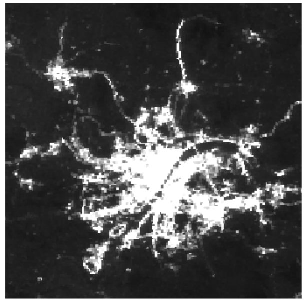
a) DNB



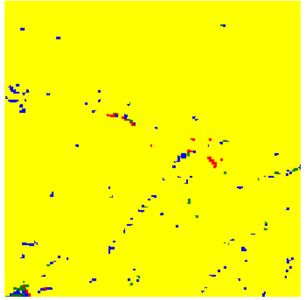
b) Cloud Mask

Day 1

Relatively cloud free



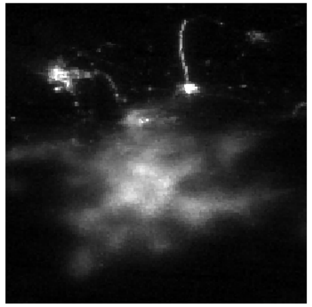
a) DNB



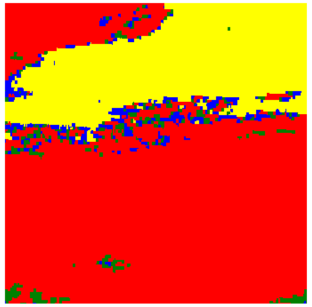
b) Cloud Mask

Day 2

Relatively cloud free



a) DNB



b) Cloud Mask

Day 3

Partly cloudy

https://blackmarble.gsfc.nasa.gov

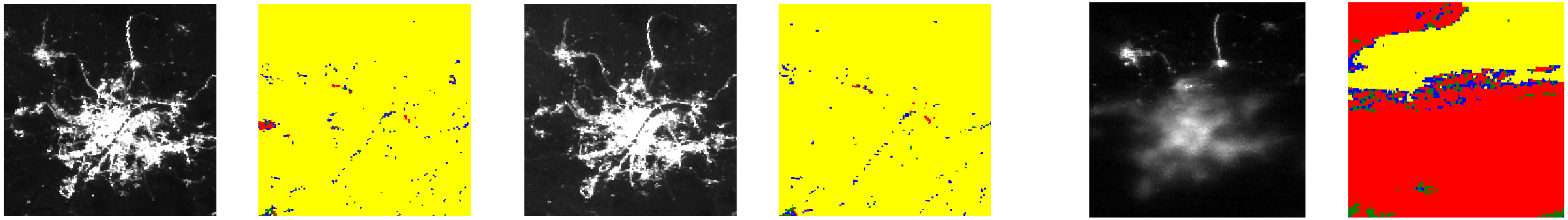
Confident clear

Probably clear

Probably cloudy

Confident cloudy

Impact of Data Quality



a) DNB b) Cloud Mask a) DNB b) Cloud Mask a) DNB b) Cloud Mask

Day 1 Day 2 Day 3

Relatively cloud free Relatively cloud free Partly cloudy

https://blackmarble.gsfc.nasa.gov

Confident clear | Probably clear | Probably cloudy | Confident cloudy

Role of Machine Learning in Determining Data Quality

- **Cloud mask**
  -detection within coral reef classification framework (Segal-Rozenhaimer et. al. 2020)
  - thin cloud removal (Singh et. al 2018)
- **Radiative Transfer Models**
  - forward (Bue et. al. 2019), inverse (ozone profile (Xu et. al. 2017), SO_2 plume height (Hedelt et. al. 2019))
  - atmospheric science (composition, trace gases)

Models and Training

Model/ approach/ algorithm selected based on question (data) at hand – unsupervised, supervised, semi-supervised

Labels

- Existing maps
- Labels from domain experts
- Discover structure from data

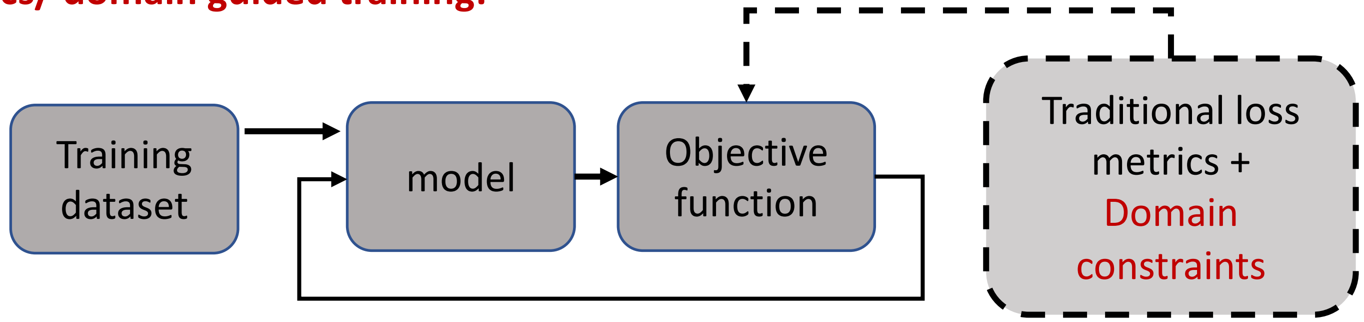
Models and Training

Model/ approach/ algorithm selected based on question (data) at hand – unsupervised, supervised, semi-supervised

Labels

- Existing maps
- Labels from domain experts
- Discover structure from data

Physics/ domain guided training:



Faghmous et. al. 2014

Ganguly et. al. 2014

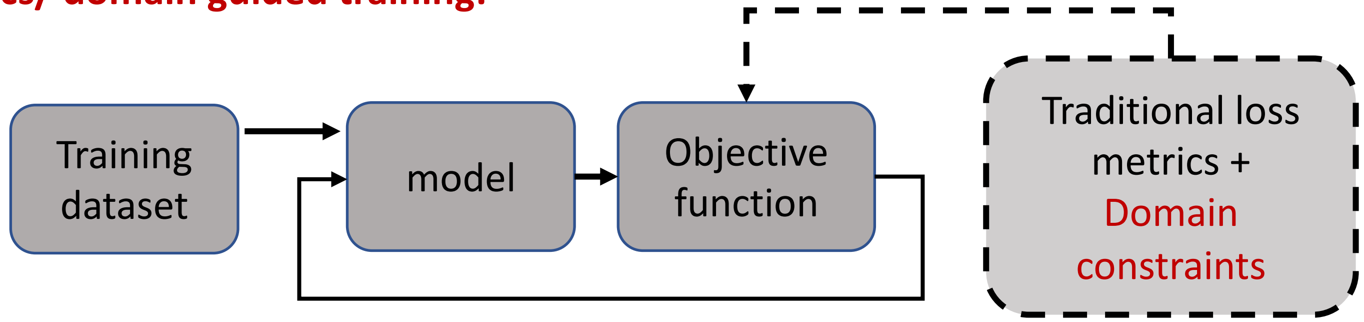
Models and Training

Model/ approach/ algorithm selected based on question (data) at hand – unsupervised, supervised, semi-supervised

Labels

- Existing maps
- Labels from domain experts
- Discover structure from data

Physics/ domain guided training:

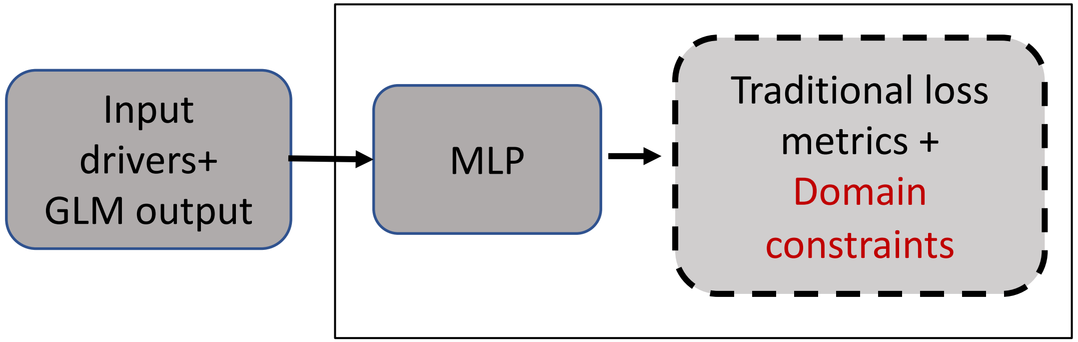


Faghmous et. al. 2014

Ganguly et. al. 2014

- **Lake Temperature:** 1-d model of lake water temperature at a given depth

- domain (physics) General Lake Models (GLM): parametrization, requires calibration, interpretable
- hybrid models of physics and data (Karpatne et. al. , 2017)
- domain relationships: f (temperature-density); density(ρ)-depth(d), $f(\rho[d_1]) - f(\rho[d_2]) \leq 0$, if $d_1 < d_2$
- loss: empirical loss + physics-guided loss; predicted temperature consistent with other physical variables

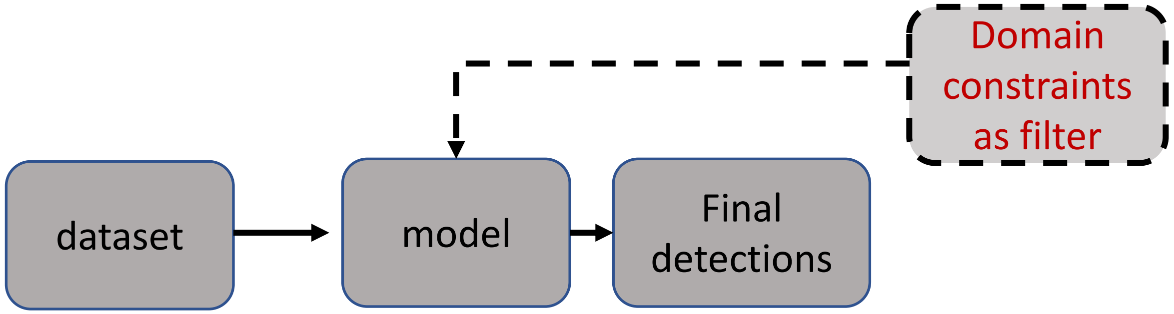


- *Domain knowledge as model input lowers prediction error, physical inconsistency*
- *Addition of loss further lowers these metrics*
- *ML+ guided model lower prediction error than lake models*

Physics-guided RNN, Jia et. al, 2019

Models and Training

Physics/ domain guide as constraints:



- **EddyScan: A physically consistent ocean eddy monitoring application**

  sea surface height from satellite altimeter AVISO data (Faghmous et. al. 2012)

- Positive and negative anomalies corresponding to eddies
- Gradual thresholding of anomalies from connected components
- Constraint by domain knowledge
  - quadratic function of eddy size(radius) vs latitude
  - isolating adjacent eddies using smallest convex hulls constraints
  - global mapping
  - empirical studies, theories
  - tracking (Faghmous et. al. 2015)

- **Remote sensing: constraints, thresholds, distributions, feature selection, estimating relationships from multivariate models**

Evaluation

Role of metrics

- ‘*Little change in global drought over the past 60 years*’, Sheffield et. al., 2012
- ‘*Increasing drought under global warming in observations and models*’, Dai, 2013
- disparate results due to varied metrics – Trenberth et. al., 2013
- long term trends from remote sensing datasets (metrics, variables)

Evaluation

Role of metrics

- ‘*Little change in global drought over the past 60 years*’, Sheffield et. al., 2012
- ‘*Increasing drought under global warming in observations and models*’, Dai, 2013
- disparate results due to varied metrics – Trenberth et. al., 2013
- long term trends from remote sensing datasets (metrics, variables)

Rare class problems

- imbalance in datasets

Evaluation

Role of metrics

- ‘*Little change in global drought over the past 60 years*’, Sheffield et. al., 2012
- ‘*Increasing drought under global warming in observations and models*’, Dai, 2013
- disparate results due to varied metrics – Trenberth et. al., 2013
- long term trends from remote sensing datasets (metrics, variables)

Rare class problems

- imbalance in datasets

Agreement with experts

- success of detection rate (detections that overlap with at least 1 expert), excess of detection rate (only detected by the method)

Chaigneau *et al*. (2008), Faghmous et. al. (2015)

Evaluation

Role of metrics

- ‘*Little change in global drought over the past 60 years*’, Sheffield et. al., 2012
- ‘*Increasing drought under global warming in observations and models*’, Dai, 2013
- disparate results due to varied metrics – Trenberth et. al., 2013
- long term trends from remote sensing datasets (metrics, variables)

Rare class problems

- imbalance in datasets

Agreement with experts

- success of detection rate (detections that overlap with at least 1 expert), excess of detection rate (only detected by the method)

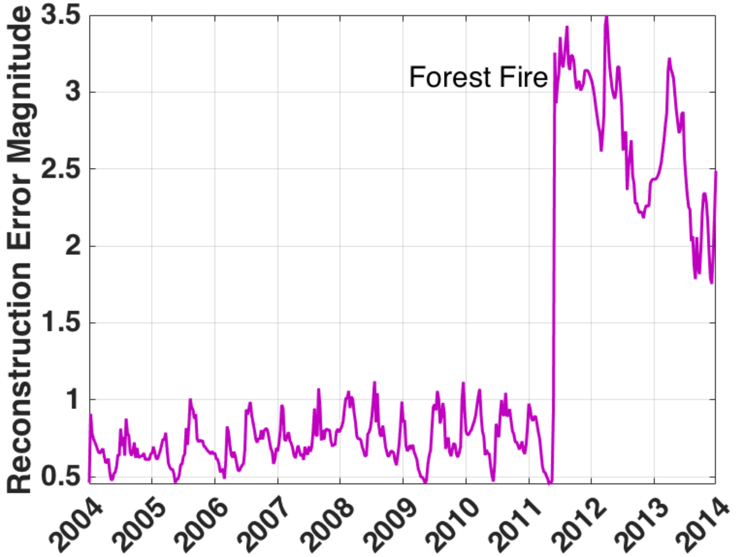
Chaigneau *et al*. (2008), Faghmous et. al. (2015)

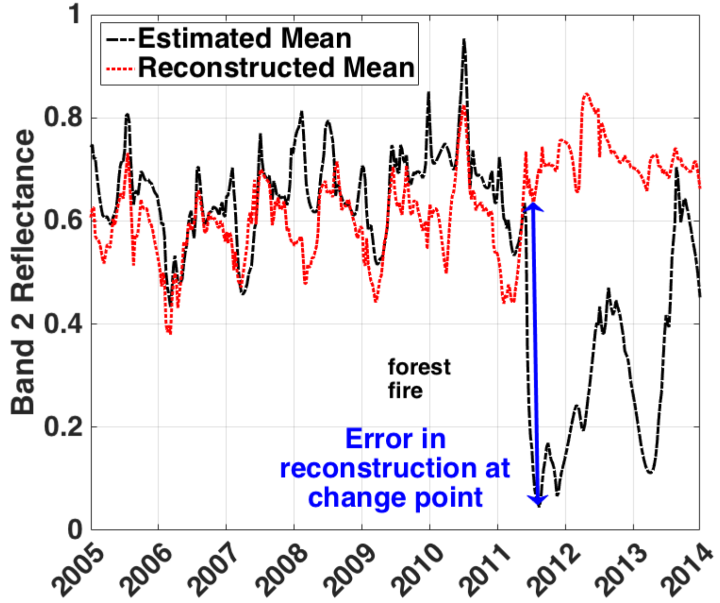
Physical Consistency

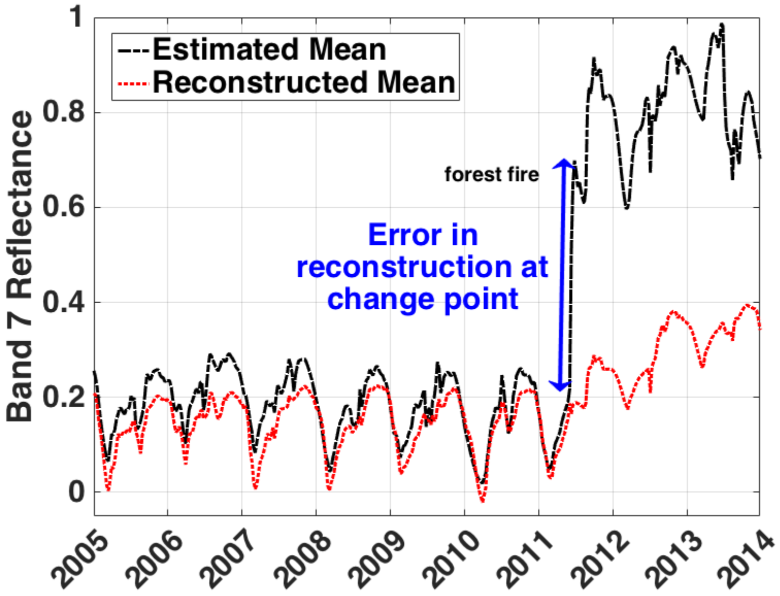
- agreement with physics-based models
- genetic algorithms (Vegetation-climate: historical NDVI, precipitation(TRMM), temperature → NDVI), Kodali et. al. 2015, Das et. al. 2016
- explanations: Land Cover Change and Natural Hazards: Model Selection and agreement with known domain science
  Chakraborty et. al., 2019a, Chakraborty 2019b

Evaluation: Explanations

Multispectral models of land surface reflectance - MODIS







Detection

Explanations: NIR

SWIR

- **Model captures domain knowledge**
- **Model selection criterion**

Chakraborty, S., 2019b, Chakraborty et. al. 2019a

Evaluation

Role of metrics

- '*Little change in global drought over the past 60 years*', Sheffield et. al., 2012
- '*Increasing drought under global warming in observations and models*', Dai, 2013
- disparate results due to varied metrics – Trenberth et. al., 2013
- long term trends from remote sensing datasets (metrics, variables)

Rare class problems

- imbalance in datasets

Agreement with experts

- success of detection rate (detections that overlap with at least 1 expert), excess of detection rate (only detected by the method)

Chaigneau *et al*. (2008), Faghmous et. al. (2015)

Physical Consistency

- agreement with physics-based models
- genetic algorithms (Vegetation-climate: historical NDVI, precipitation(TRMM), temperature → NDVI), Kodali et. al. 2015, Das et. al. 2016
- explanations: Land Cover Change and Natural Hazards: Model Selection and agreement with known domain science
  Chakraborty et. al., 2019a, Chakraborty 2019b

Energy Consumption

- emissions from machine learning – *Machine Learning Emissions Calculator,* Lacoste et. al. 2019

Domain Expert Feedback

Prepare data	Training	Evaluation	Inference/ Feedback
- Problem framing - Dataset selection - Provide labels	-Domain knowledge -Constraints -Distributions -Features, weights	-Assess performance -Satisfies expected physical laws?	-Acceptability -Retrain, update -Feedback on misclassifications

- **Disagreement between model and experts**
- **Disagreement between experts**
- **Expert effort**
- **Explanation of decisions**

Conclusion

Data Quality	Model and Training	Evaluation	Domain Expert Feedback
- Sample selection - Role in detecting data quality	- Couple datasets with domain guides for consistency	- Appropriate metrics - Explanation - Climate impact	- Validation - Labels - Training - acceptability

Additional challenges:

- Integrating heterogeneous data - representations
- Spatio-temporal heterogeneity
- Adaptive models
- Impact of machine learning

References

1. Segal-Rozenhaimer, M., Li, A., Das, K. and Chirayath, V., 2020. Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN). *Remote Sensing of Environment*, *237*, p.111446.
2. Xu, J., Schüssler, O., Rodriguez, D.G.L., Romahn, F. and Doicu, A., 2017. A novel ozone profile shape retrieval using full-physics inverse learning machine (FP-ILM). *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(12), pp.5442-5457.
3. Hedelt, P., Efremenko, D., Loyola, D., Spurr, R. and Clarisse, L., 2019, January. SO2 Layer Height retrieval from Sentinel-5 Precursor/TROPOMI using FP-ILM. In *Geophysical Research Abstracts* (Vol. 21).
4. Bue, B.D., Thompson, D.R., Deshpande, S., Eastwood, M., Green, R.O., Natraj, V., Mullen, T. and Parente, M., 2019. Neural network radiative transfer for imaging spectroscopy. *Atmospheric Measurement Techniques*, *12*(4).
5. Singh, P. and Komodakis, N., 2018, July. Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*(pp. 1772-1775). IEEE.
6. Martinez, E., Gorgues, T., Lengaigne, M., Fontana, C., Sauzède, R., Menkes, C., Uitz, J., Di Lorenzo, E. and Fablet, R., 2020. Reconstructing Global Chlorophyll-a Variations Using a Non-linear Statistical Approach. *Frontiers in Marine Science*, *7*, p.464.
7. Faghmous, J.H. and Kumar, V., 2014. A big data guide to understanding climate change: The case for theory-guided data science. *Big data*, *2*(3), pp.155-163.
8. Ganguly, A.R., Kodra, E.A., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S., Chatterjee, S., Choudhary, A., Das, D., Faghmous, J. and Ganguli, P., 2014. Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques.
9. Karpatne, A., Watkins, W., Read, J. and Kumar, V., 2017. Physics-guided neural networks (pgnn): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*.
10. Jia, X., Willard, J., Karpatne, A., Read, J., Zwart, J., Steinbach, M. and Kumar, V., 2019, May. Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 558-566). Society for Industrial and Applied Mathematics.
11. Faghmous, J.H., Styles, L., Mithal, V., Boriah, S., Liess, S., Kumar, V., Vikebø, F. and dos Santos Mesquita, M., 2012, October. Eddyscan: A physically consistent ocean eddy monitoring application. In *2012 Conference on Intelligent Data Understanding* (pp. 96-103). IEEE.
12. Faghmous, J. H. et al. Multiple hypothesis object tracking for unsupervised self-learning: An ocean eddy tracking application. In *Proc. 27th AAAI Conf. Artif. Intel.* (2013).
13. Faghmous, J.H., Frenger, I., Yao, Y., Warmka, R., Lindell, A. and Kumar, V., 2015. A daily global mesoscale ocean eddy dataset from satellite altimetry. *Scientific data*, *2*(1), pp.1-16.
14. Sheffield, J., Wood, E.F. and Roderick, M.L., 2012. Little change in global drought over the past 60 years. *Nature*, *491*(7424), pp.435-438.
15. Dai, A., 2013. Increasing drought under global warming in observations and models. *Nature climate change*, *3*(1), pp.52-58.
16. Trenberth, K.E., Dai, A., Van Der Schrier, G., Jones, P.D., Barichivich, J., Briffa, K.R. and Sheffield, J., 2014. Global warming and changes in drought. *Nature Climate Change*, *4*(1), pp.17-22.
17. Chaigneau, A., Gizolme, A. and Grados, C., 2008. Mesoscale eddies off Peru in altimeter records: Identification algorithms and eddy spatio-temporal patterns. *Progress in Oceanography*, *79*(2-4), pp.106-119.
18. Das, K., Kodali, A., Szubert, M., Ganguly, S. and Bongard, J., 2016. Quantifying How Climate Affects Vegetation in the Amazon Rainforest. *AGUFM*, *2016*, pp.IN11B-1621.
19. Kodali, A., Khandelwal, A., Ganguly, S., Bongard, J. and Das, K., 2015. Regression based modeling of vegetation and climate variables for the Amazon rainforests. *AGUFM*, *2015*, pp.IN51A-1799.
20. Chakraborty, S., Papandreou-Suppappola, A. and Christensen, P.R., 2019a. Class Separability of Land Cover Change Events from Multispectral Satellite Image Time-Series. *AGUFM*, *2019*, pp.IN43A-05.
21. Chakraborty, S., 2019b. *Advanced Processing of Multispectral Satellite Data for Detecting and Learning Knowledge-based Features of Planetary Surface Anomalies* (Doctoral dissertation, ASU).
22. Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T., 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.