Cornell University

Innovations in Feed the Future Monitoring and Evaluation: Harnessing Big Data and Machine Learning to Feed the Future

Cornell University team

Presentation to USAID

March 30, 2021

Cornell University

Innovations in Feed the Future Monitoring and Evaluation: Harnessing Big Data and Machine Learning to Feed the Future

Project Team:
Cornell: Jiaming Wen, Ying Sun, Linden McBride, David Matteson, Oz Kira, Medha Bulumulla, Chris Browne, Chris Barrett
Alabama: Leiqiu Hu
IFPRI: Yanyan Liu, Susana Constenla

USAID AOR: Farzana Ramzan (BRFS)

Cooperative Agreement #7200AA18CA00014 b/n USAID and Cornell

Meeting Objectives

1) Update USAID on project findings, with an eye to implications for USAID and its partners' strategy(ies) for the next few years

2) Agree on final project stage outputs, incl. training/outreach given results and COVID restrictions.

Meeting Agenda

1. Introductions
2. Project overview (Chris Barrett)
3. Explain LST data products (Leiqiu Hu)
4. Explain SIF data products (Ying Sun)
5. Maize yield prediction w/SIF (Oz Kira and Ying Sun)
6. Multivariate random forest prediction of poverty and malnutrition outcomes (Chris Browne and David Matteson)
7. Sequential prediction using monthly NDMA data from Kenya (Yanyan Liu and Chris Barrett)
8. Big data and machine learning for prediction of human well-being indicators (Linden McBride)
9. General discussion

Project Overview

Our project objective: Reasonably accurate, low-cost, timely prediction of human well-being (FtF) indicators for rural communities are essential for early warning, targeting, and MEL purposes. But still too hard and expensive to develop.

We developed new open source data products (SIF, LST), link those w/other open access, near-real-time data sources in simple ML-based prediction of crop yield, poverty, nutrition indicators at ADM3(or higher)-level spatial resolution. Can cheaper, higher frequency, operationally implementable estimates be sufficiently accurate to supplement (or fill the void from the absence of) large-scale survey-based measures generated infrequently? For what tasks might this work?

New monthly LST products with improved sensitivity to drought

Long-term records (2010-2018) over the FtF countries

High resolution at 1km

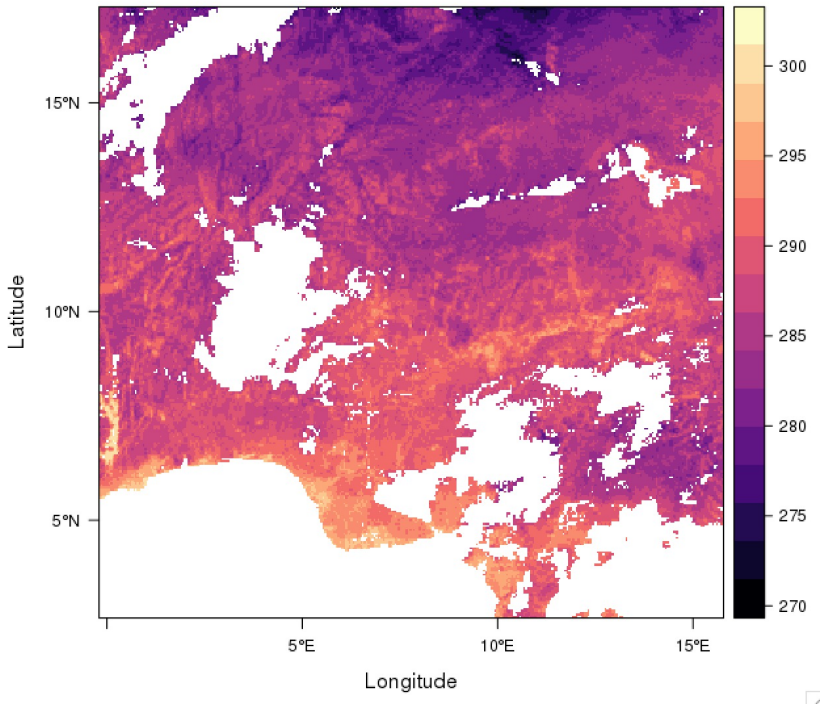
Temporal and spatial Gap-filled data that improve the spatial consistency and temporal representation

Leiqiu Hu, University of Alabama in Huntsville

Overview (Motivations)

- Remotely sensed LST provides spatially explicit and temporally frequent information linked with vegetation stress associated with droughts.
- To accurately detect the temperature anomalies and to pinpoint the impacted regions, spatially-complete and temporally representative LST data are needed.
- This project supports development of 1-km monthly LST composite:
  - 24hr (daily) mean LST
  - mean diurnal LST range
  - daily maximum and minimum LST.

- Missing data due to clouds
- Hourly products with coarser spatial resolution (5 km)



Nigeria: 2013-01-05 4:00 UTC (SEVIRI)

Modeling framework

GEO LST
(Hourly, 5km)

MODIS LST
(Sub-daily, 1km)

DTC Model

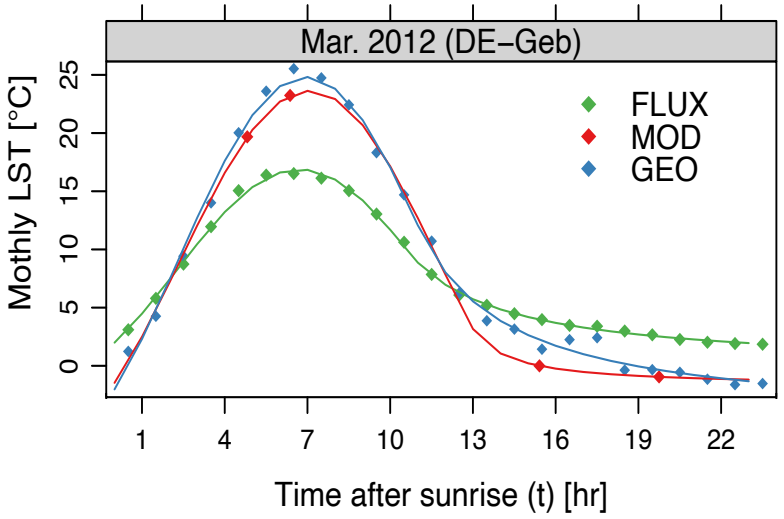
Gap-fill missing data due to clouds and infrequent observations

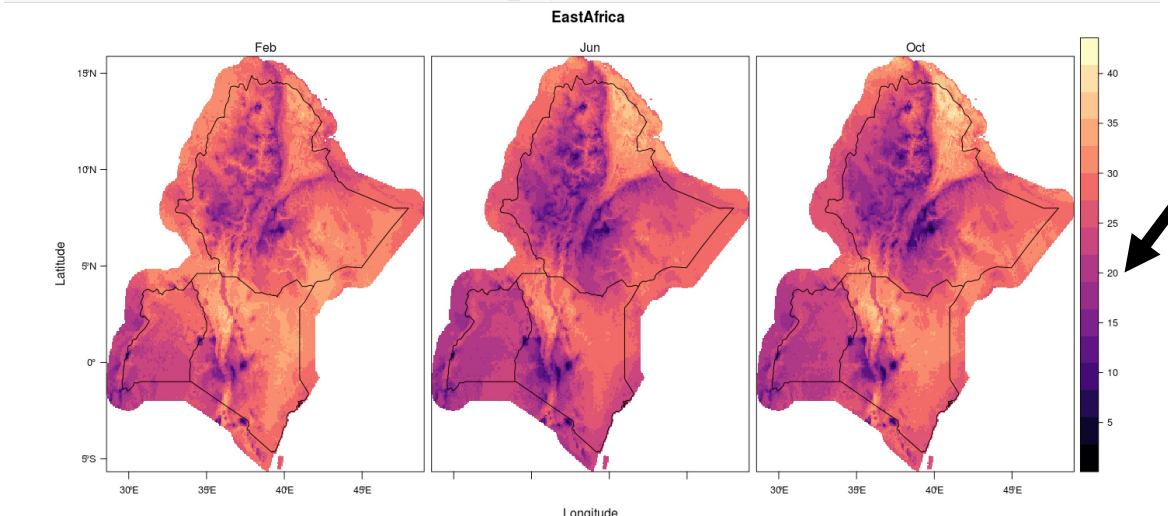
24-h mean LSTs
Daily LST extrema
(monthly, 1km)

Validation

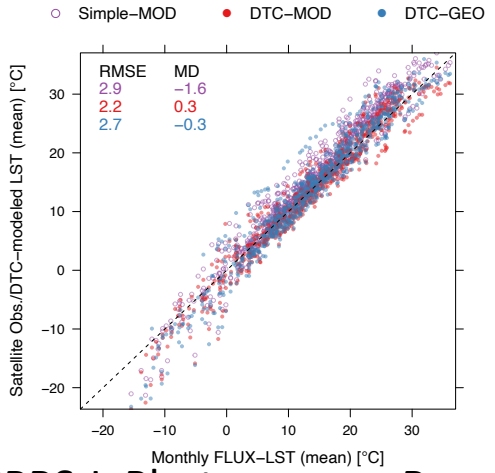
Global Flux sites

Diurnal Temperature Cycle (DTC) Model



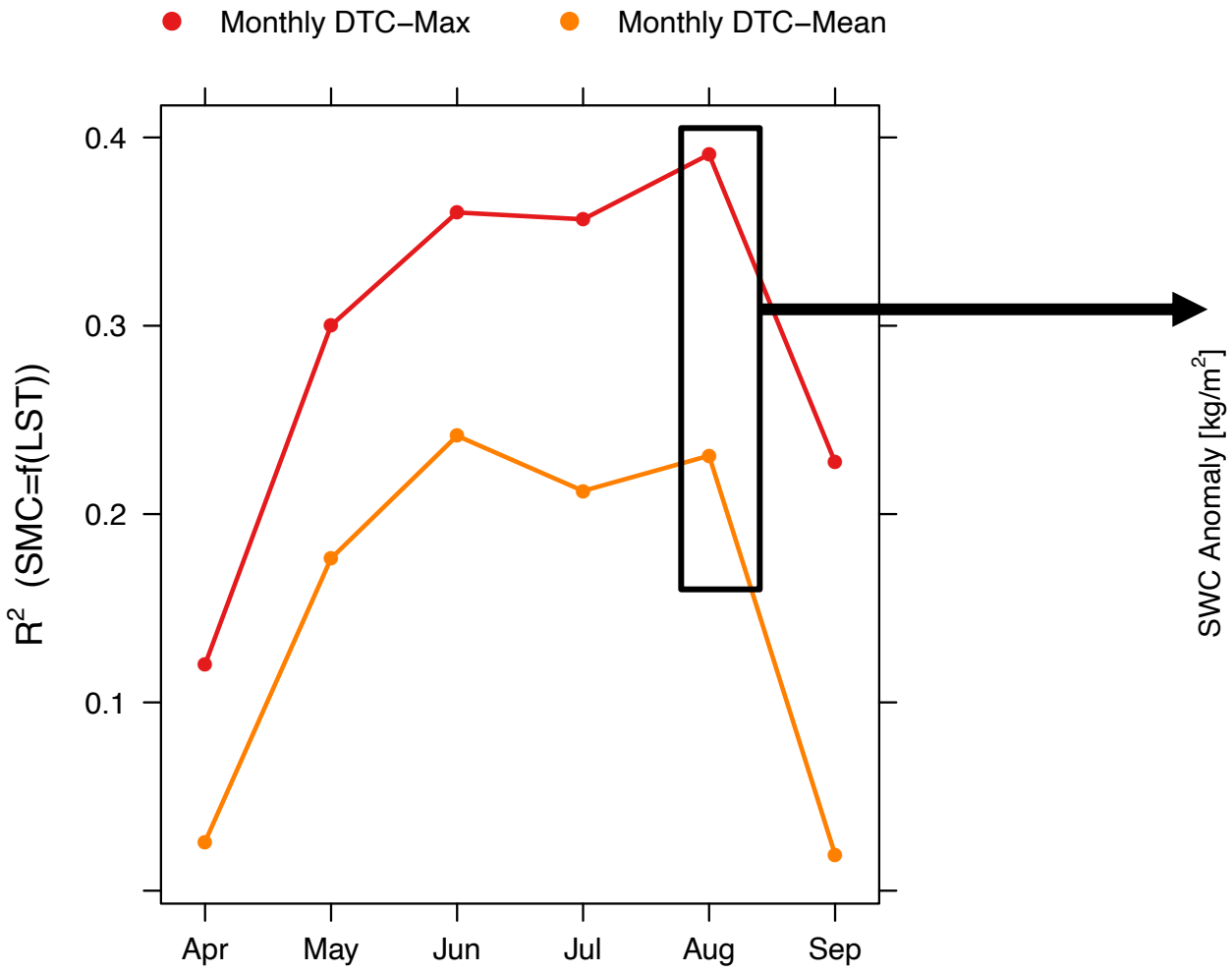


RMSE: 2.2 K
Mean diff.: 0.3 K

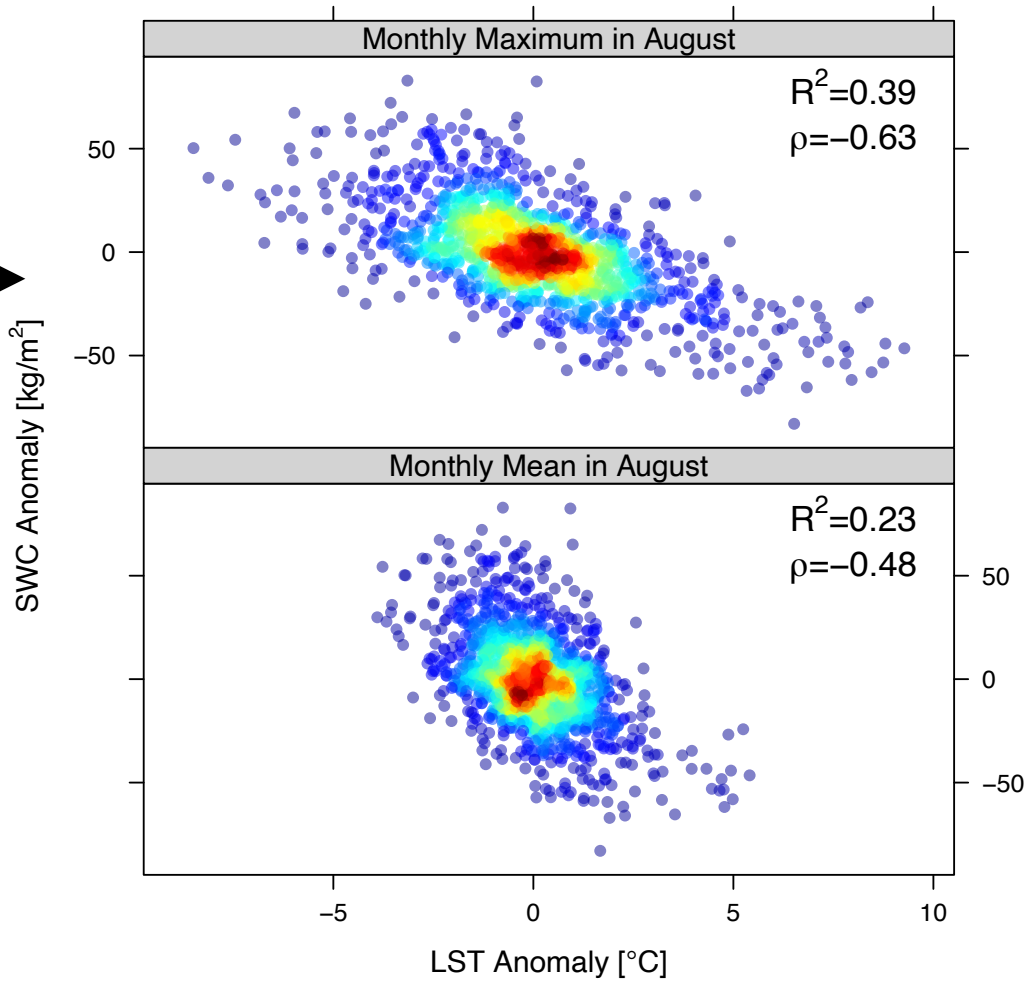


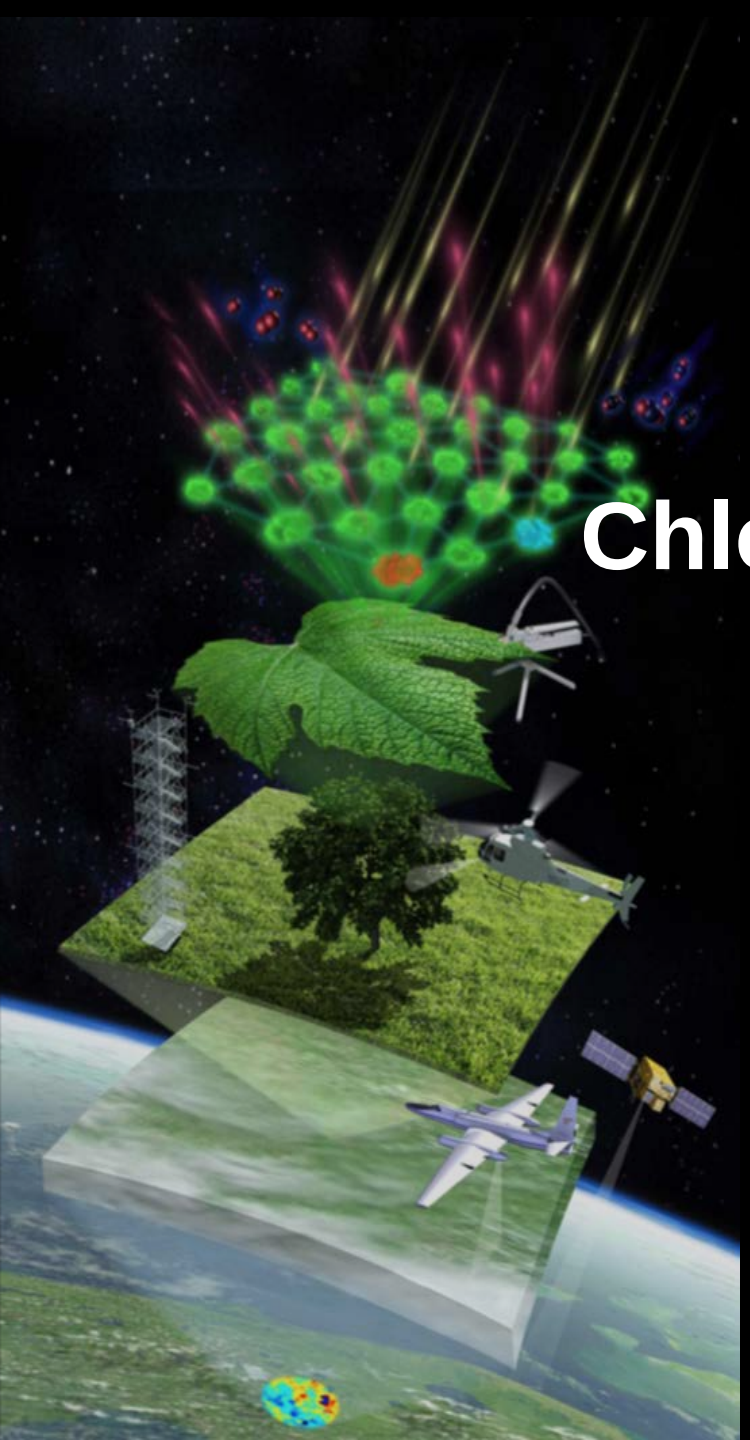
(Hu et al., 2020, ISPRS J. Photogramm. Remote Sens.)

Example: Drought sensitivity



Soil Moisture Content (SMC)



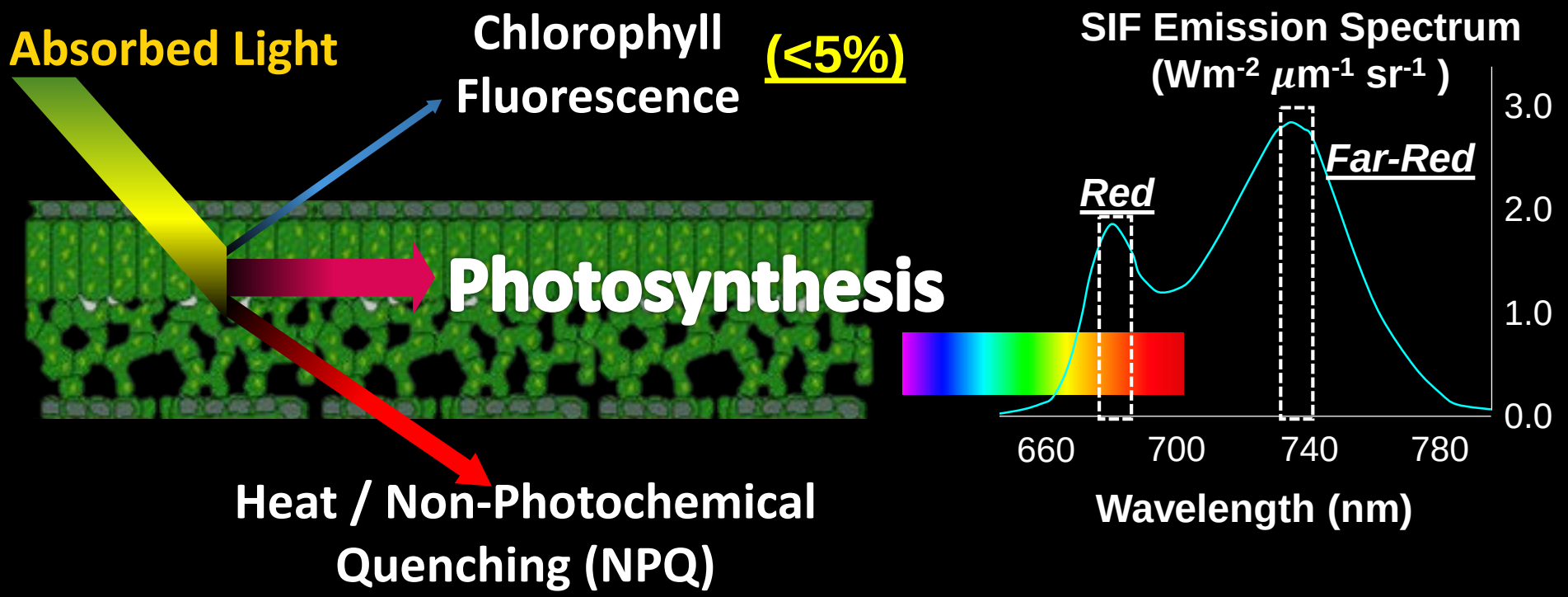


Remote Sensing of Solar-Induced Chlorophyll Fluorescence (SIF) Improves Photosynthesis Estimation

Ying Sun, Jiaming Wen

School of Integrative Plant Science
Cornell University
ys776@cornell.edu

Solar-Induced chlorophyll Fluorescence (SIF): A probe of photosynthesis *in vivo*



Direct information of plant functioning

A: Structural: How much light is absorbed by the antenna system?

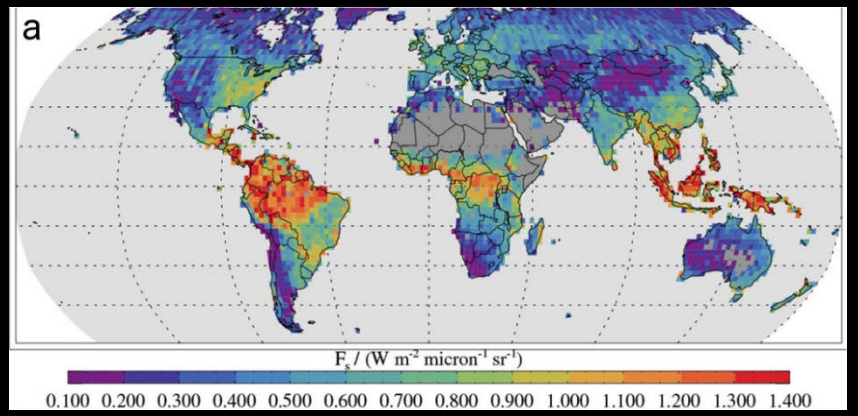
B: Physiological: How efficiently is this light used?

The emerging SIF remote sensing offers potential to revolutionize GPP measurement

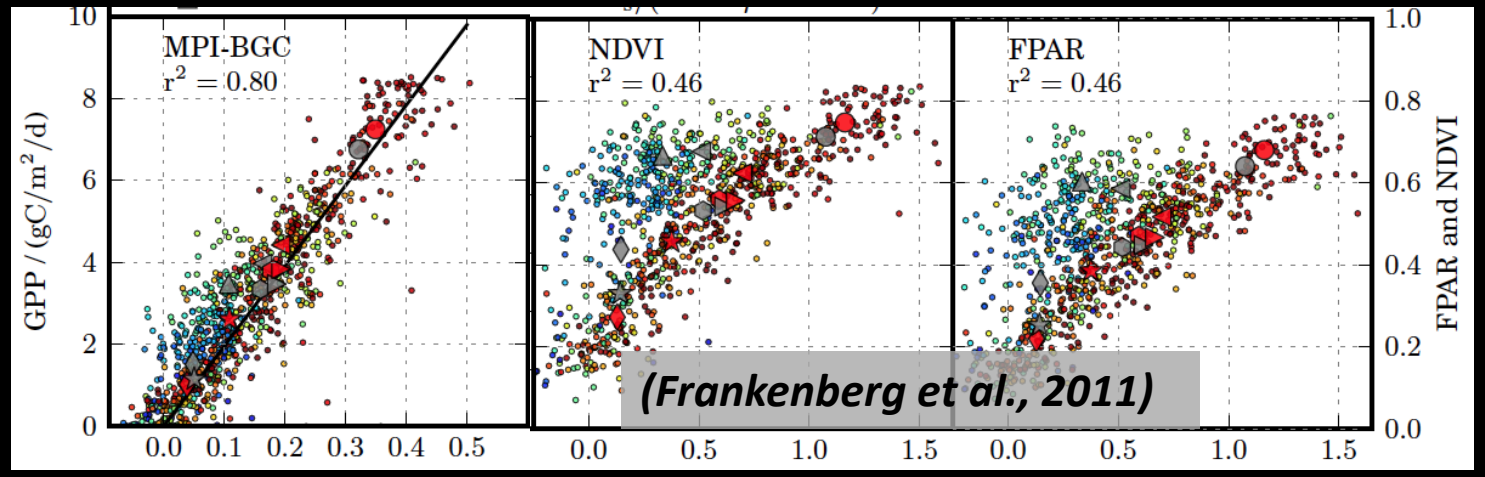
GOSAT SIF

Annual mean (June2009-May 2010)

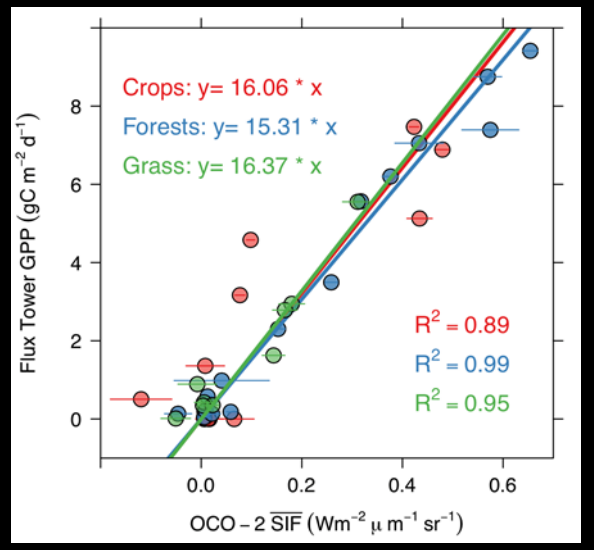
(Frankenberg et al., 2011)



- ✓ *Less sensitive to aerosols/thin cloud*
- ✓ *Functional proxy (vs empirical proxy from conventional vegetation indices-based estimates)*
- ✓ *No need for ancillary information to estimate GPP*
- ✓ *Early warning of stress*



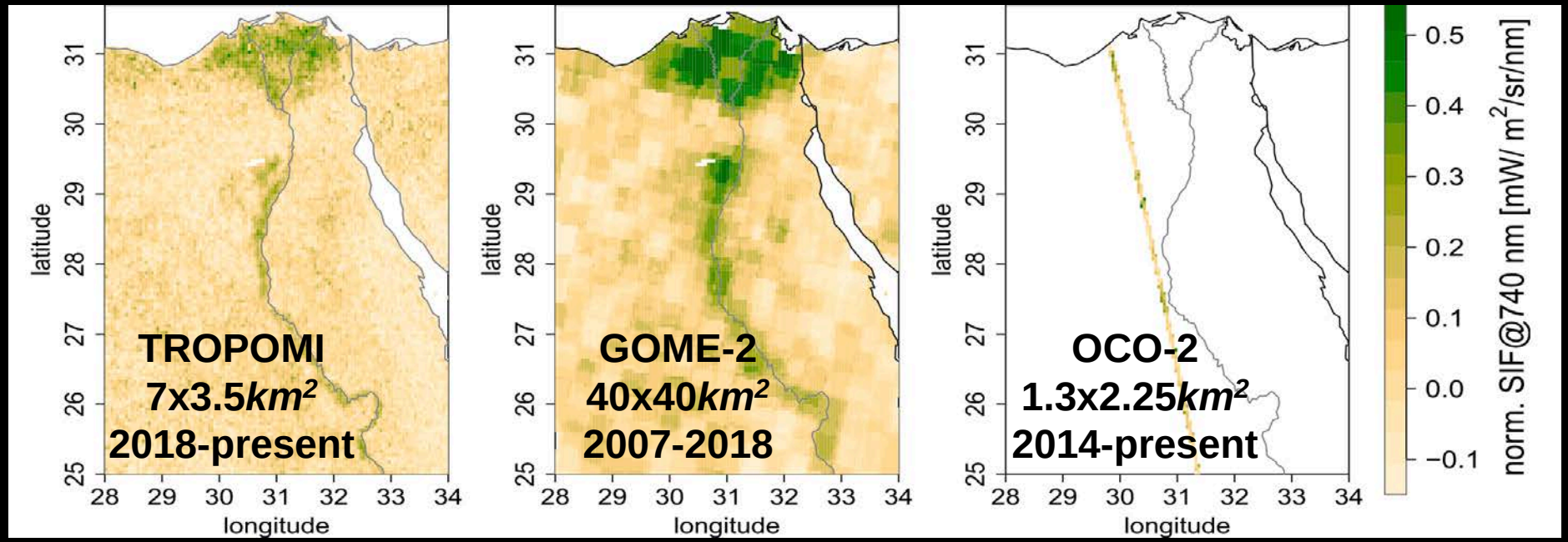
(Frankenberg et al., 2011)



(Sun et al., 2017)

Limitations of existing satellite SIF products

Short time record **Low spatial resolution** **Large spatial gaps**



Goal: Develop high-resolution, globally contiguous, long-term SIF records

Strategy:

Machine Learning (data-driven) + Physiological Constraints (process)

Gap-filling of OCO-2 SIF

MODIS surface reflectance (7-band)

+

SIF along orbits → **ML** ← +

Stratify time and biomes

SIF = PAR * fPAR * LUE

SIF at 0.05°

SIF_{OCO2_005}: *2014 - present*

Yu et al., 2019, GRL

Downscaling of GOME-2 & SCIAMACHY SIF

MODIS surface reflectance (7-band)

+

Stratify time and space → **ML** ← + SIF at 0.5°

SIF = PAR * fPAR * LUE

SIF at 0.05°

SCIA: 2002 - 2012 + GOME-2: 2007- 2018

↓ CDF matching

SIF_{005}: *2002 - 2018*

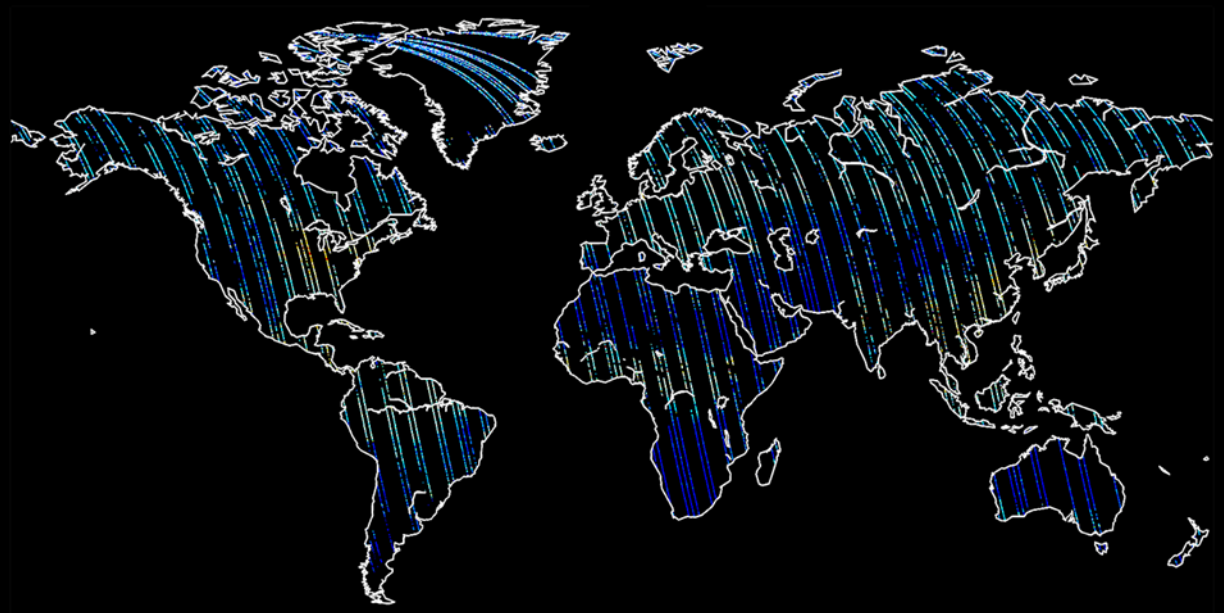
Wen et al., 2020, RSE

I: Gap-filling of OCO-2 SIF

Construct the spatial continuity

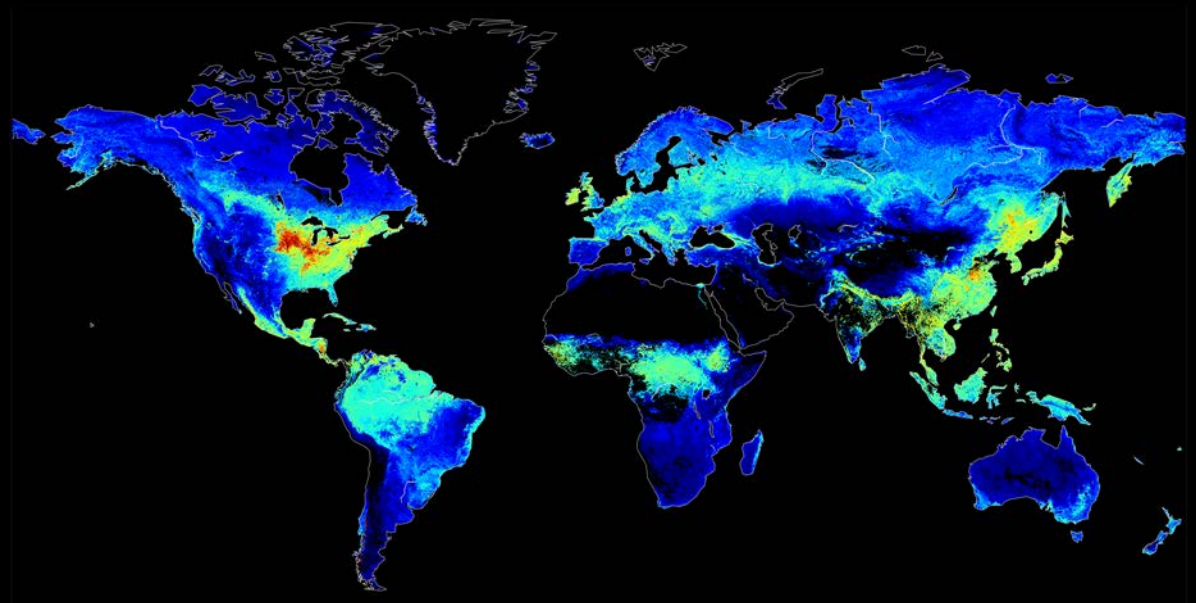
Original SIF along orbits

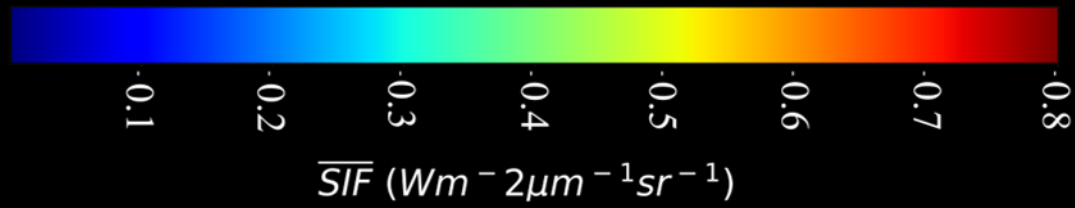
(1st 16-day of 08/2015)



Gap-filled SIF at 0.05°

(1st 16-day of 08/2015)

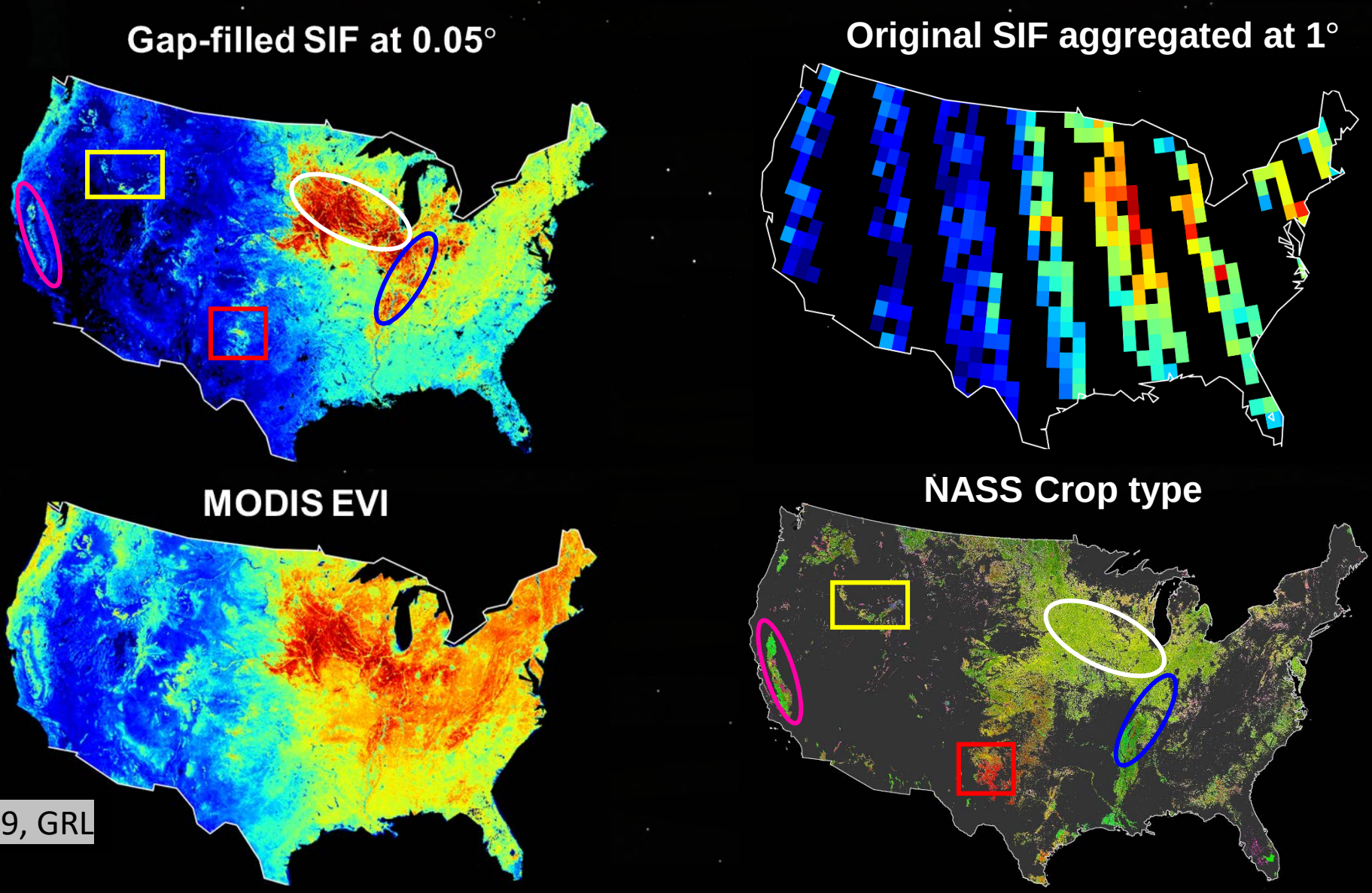




Yu et al., 2019, GRL

I: Gap-filling of OCO-2 SIF

Identify highly productive agricultural systems

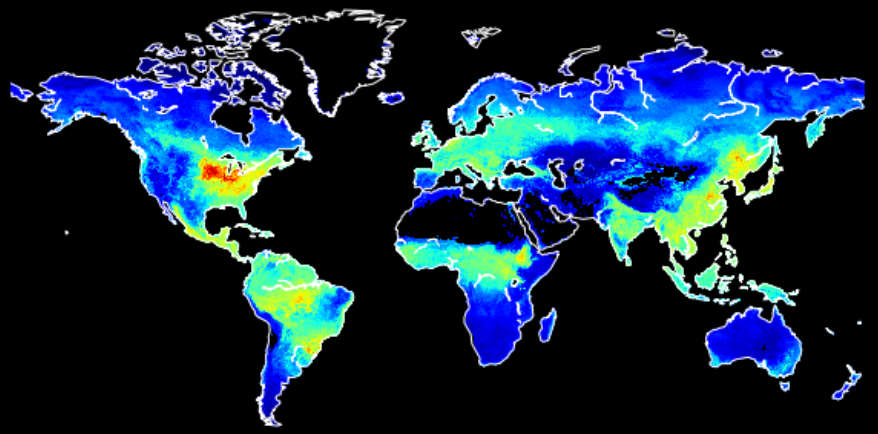


Yu et al., 2019, GRL

II: Downscaling of GOME-2 and SCIAMACHY SIF

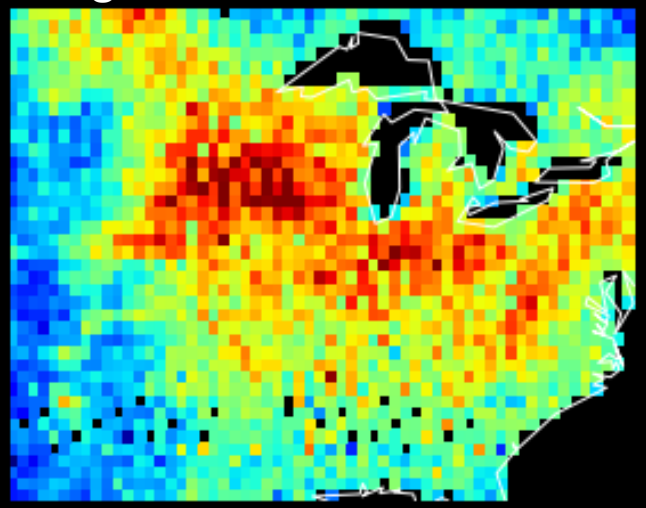
Reveal fine spatial details

SIF005 (August 2015)

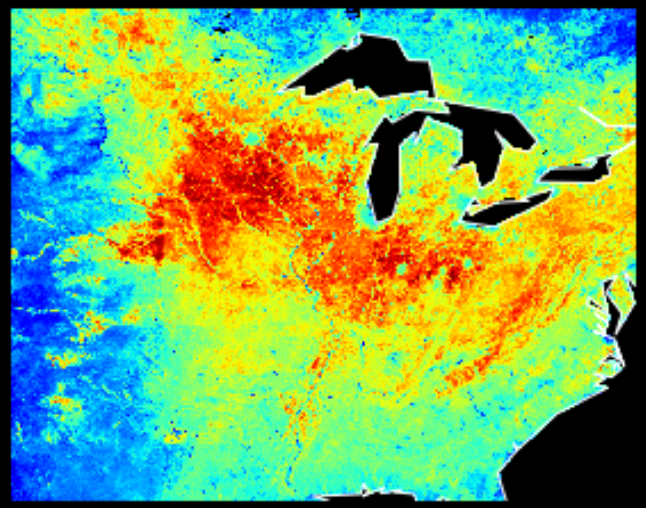


Wen et al., 2020, RSE

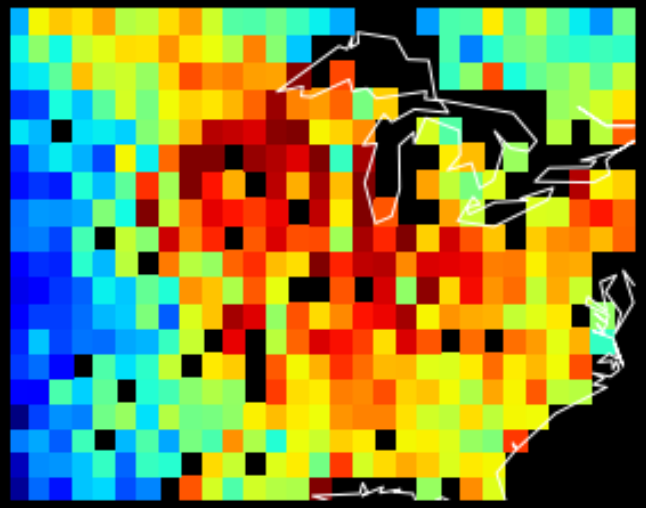
Original GOME-2 SIF at 0.5°



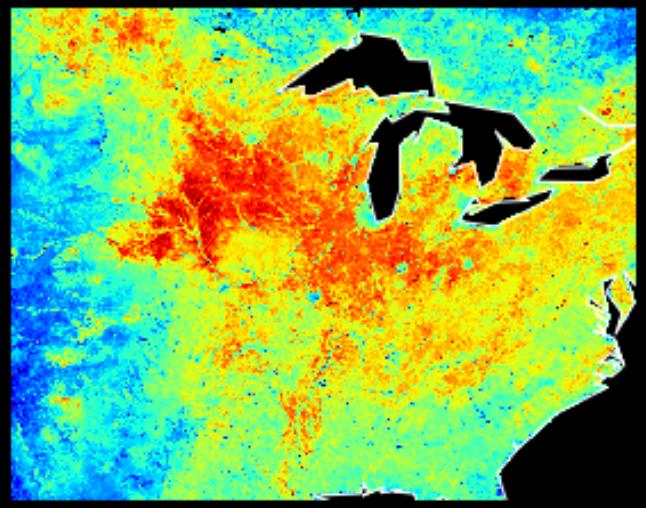
GOME-2 SIF at 0.05°

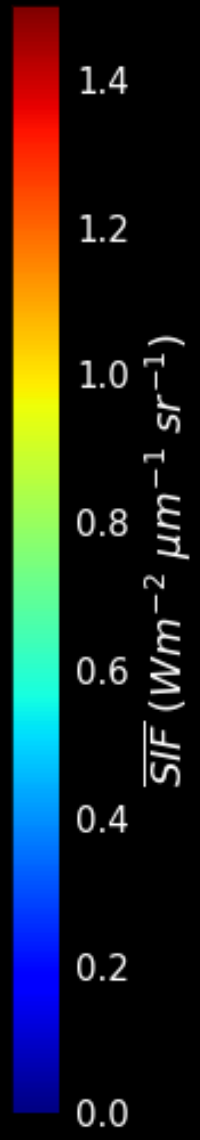


Original SCIA SIF at 1°



Scaled SCIA SIF at 0.05°





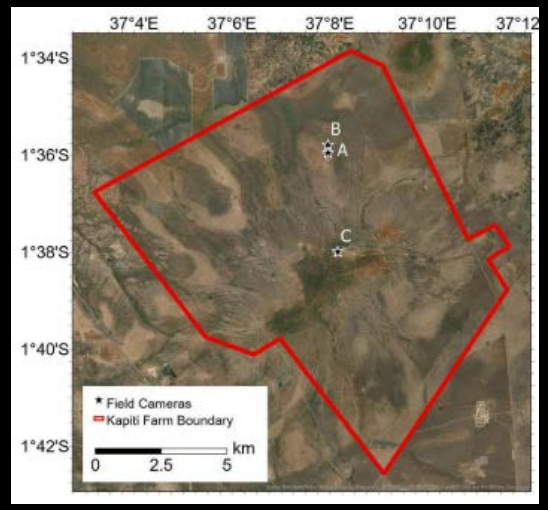
Applications

- **Quantify the CO2 uptake by terrestrial ecosystems**
- **Estimate crop yield in US Corn Belt and India**
- **Predict poverty and malnutrition indicators in food-insecure countries**
- **Evaluate land degradation in Africa**

II: Downscaling of GOME-2 and SCIAMACHY SIF

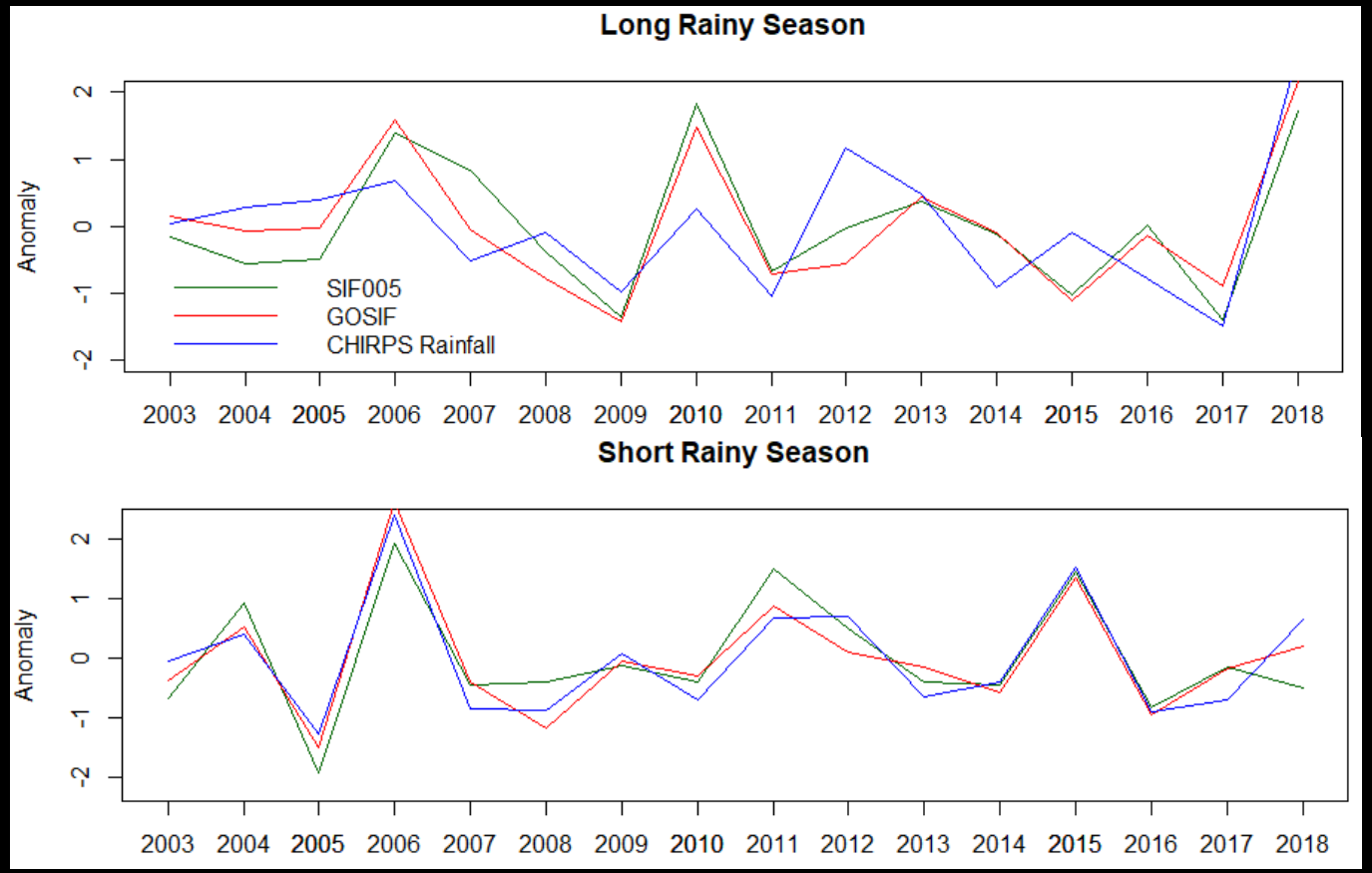
Preserves interannual variability

Kapiti farm in Kenya (grasslands)



Cheng et al., 2020

SIF005 tracks the interannual variation dominant by precipitation



CornellCALS
College of Agriculture and Life Sciences

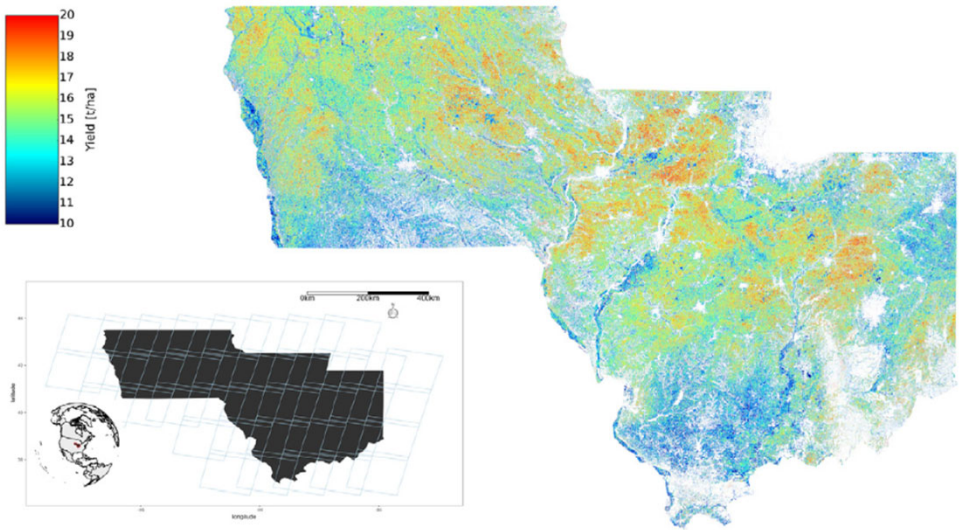
Process based crop yield model using Solar induced chlorophyll fluorescence

Oz Kira and Ying Sun

Scalable crop yield mapper

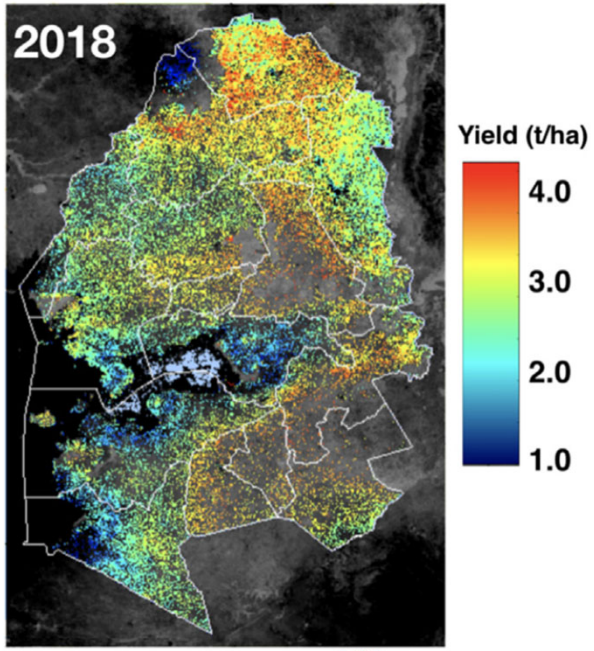
A crop yield model that does not depend on heavy calibration with ground observations, which are often not readily available and/or in good quality especially in food insecurity countries, would greatly benefit large-scale yield monitoring and early warning.
e.g., Scalable crop yield mapper (SCYM) (Lobell et al., 2015. RSE)

India



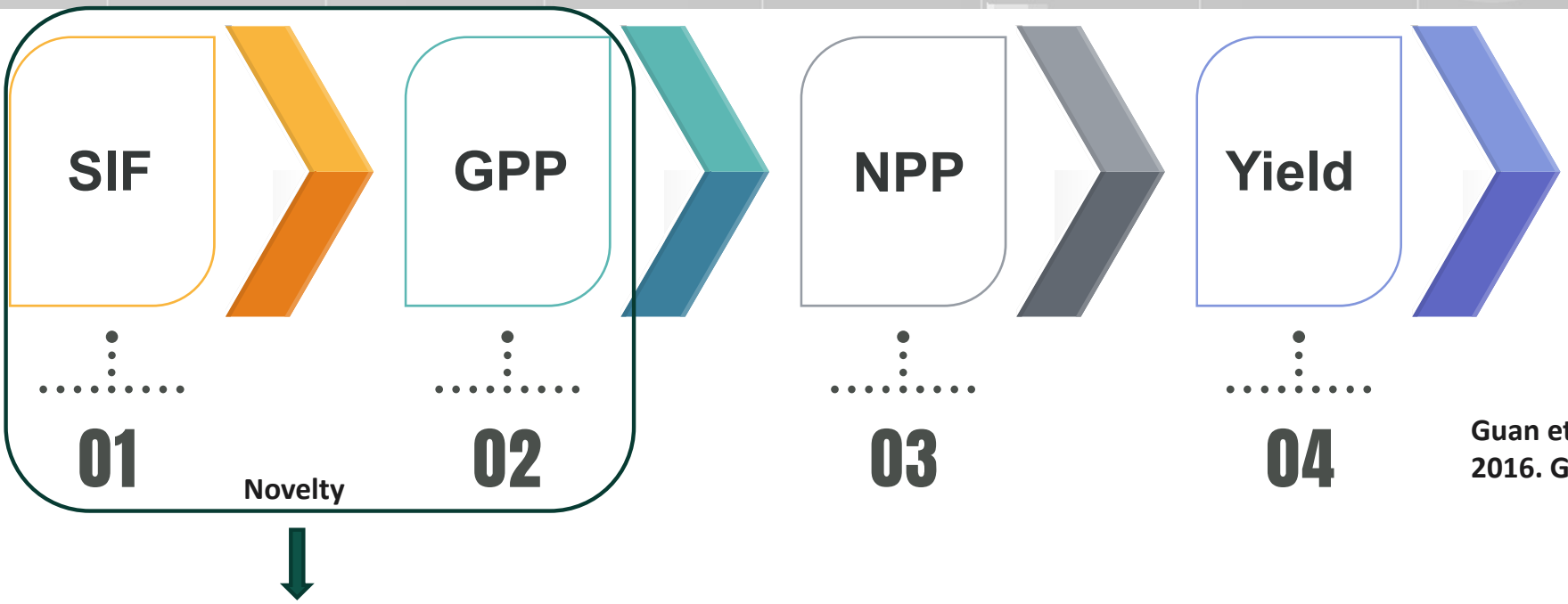
Azzari et al., 2017. RSE

Kenya and Tanzania



Jin et al., 2017. RSE

Process based yield model



Guan et al., 2016. GCB

Mechanistic light reaction model, which is capable of accurately estimating photosynthesis with SIF as an observational input by establishing the theoretical relationship between SIF and the actual electron transport rate from photosystem II to photosystem I.

Gu et al., 2019. New Phytologist

SIF to GPP

Purple – measured/observed

Red – parameters, largely conservative and well established in existing literatures

Fraction of open PSII reaction centers:

$$q_L = a_{q_L} e^{-b_{q_L} PAR}$$

Electron transport:

$$J = \frac{\Phi_{PSIImax} \cdot (1 + k_{df})}{(1 - \Phi_{PSIImax}) \cdot \varepsilon} \cdot q_L \cdot SIF_{\ unmixed}$$

Gross primary production:

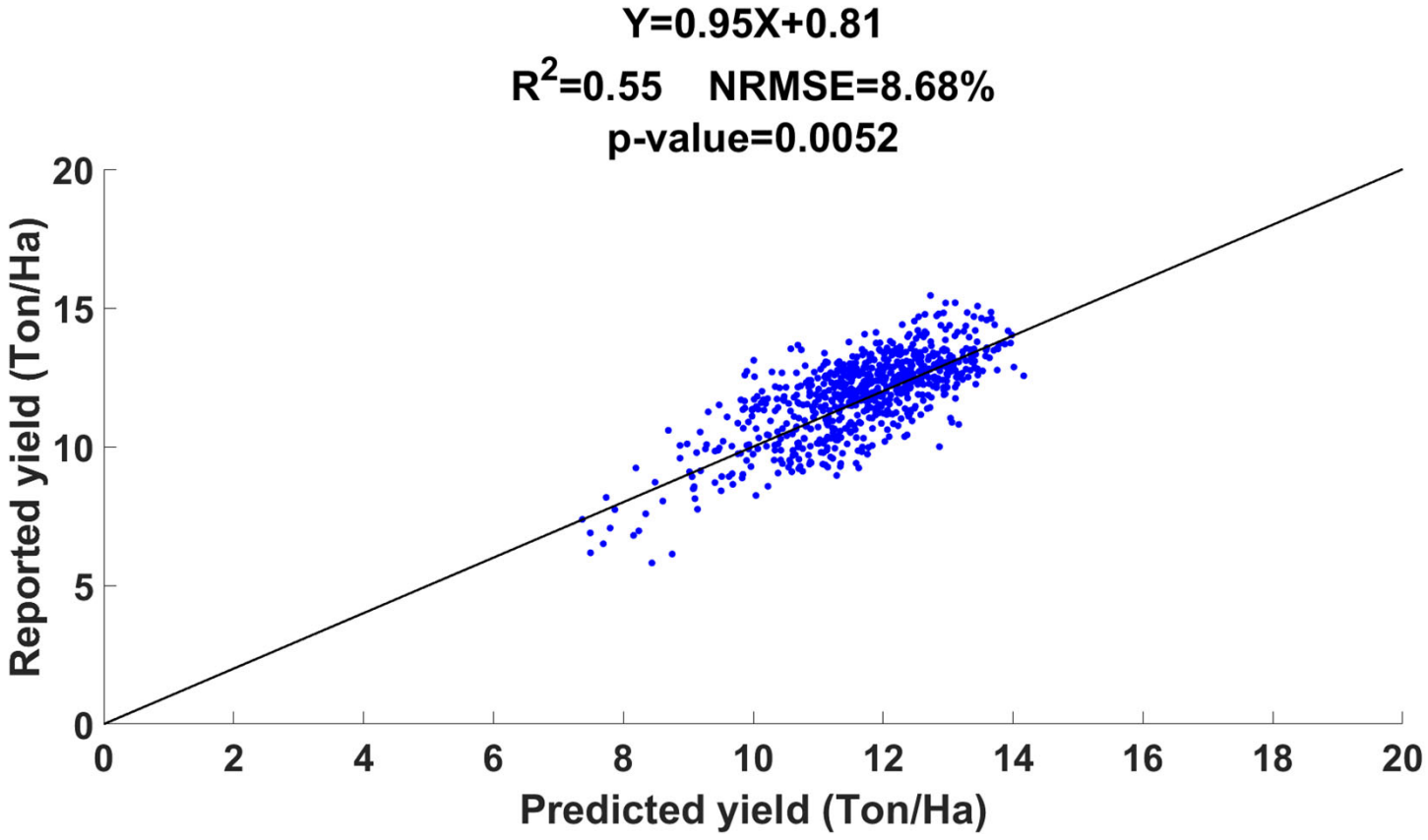
$$GPP = \frac{1 - x}{3} \cdot J$$

Produced from satellite observations: ***PAR (hourly averaged on half a month), SIF (half monthly), ε (half monthly).***

The used *SIF* is generated by an unmixing procedure

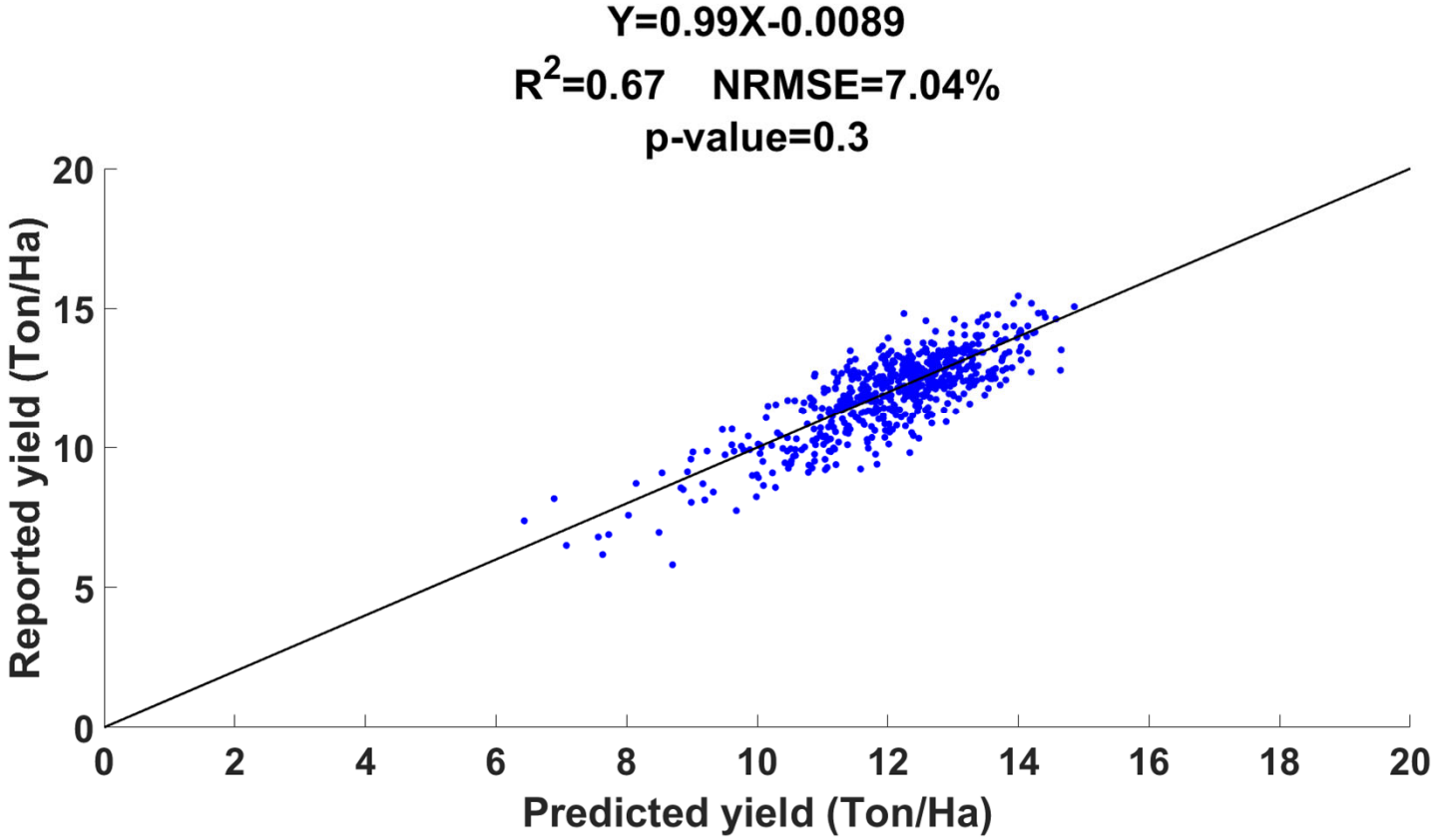
(for more details please see: Kira and Sun., Extraction of sub-pixel C3/C4 emissions of solar-induced chlorophyll fluorescence (SIF) using artificial neural network).

Process-based model (no calibration) in the US corn belt



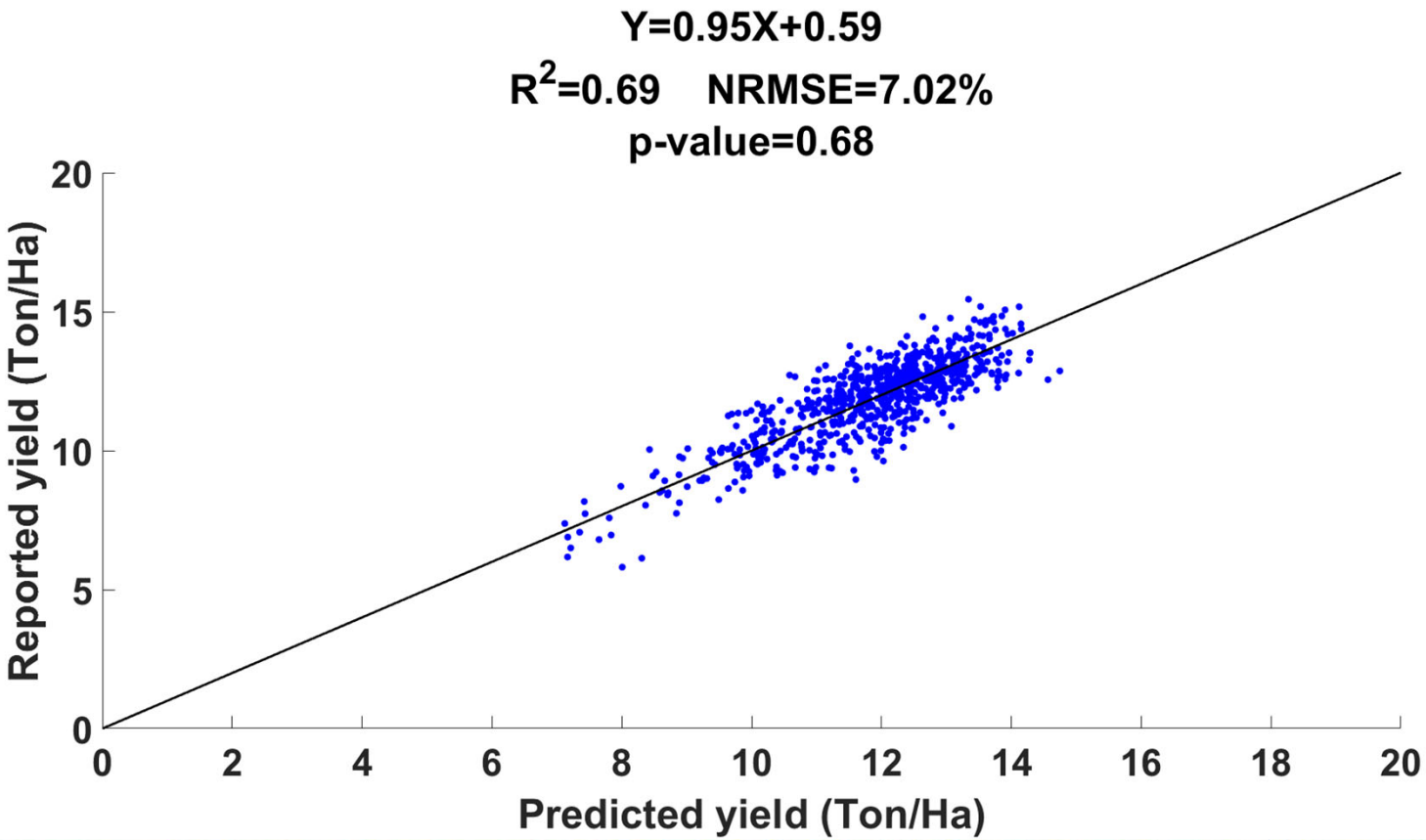
Neural network model in the US corn belt

Model based on 8 half monthly (June-September) values of SIF.



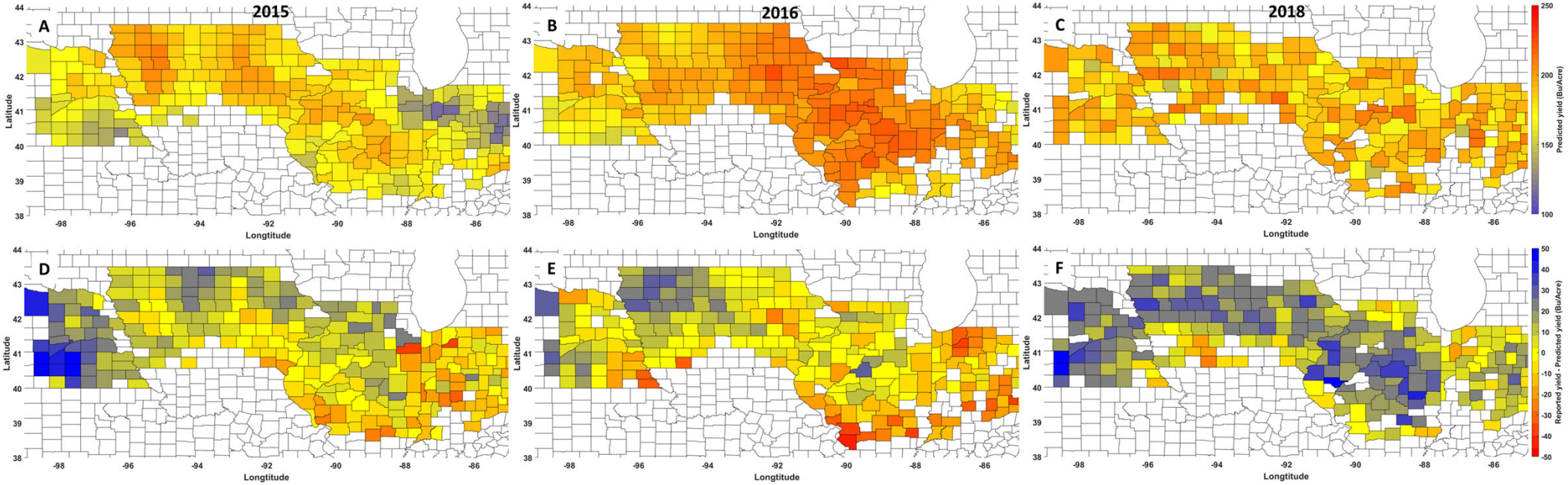
Process-based model in the US corn belt – optimization of qL model

After optimizing a_{qL} and b_{qL} - $q_L = a_{q_L} e^{-b_{q_L} PAR}$



Results of the process based model in US Corn belt

Spatial distribution of predicted yield using the un-calibrated model (A, B, C) and error of prediction (D, E, F) for 2015, 2016, and 2018. Error of prediction is defined as the difference between the reported yield and the predicted yield.



Conclusions

- **The process-based model with no calibration against yield observations performs reasonably well to capture both the magnitude and variability of corn yield in the US Midwest (where ground truth is available for verification).**
- **Fine tuning of a minimum number of parameters can further improve the model performance, which is at least as good as heavily calibrated neural network types of model.**
- **This opens up new opportunities for estimating crop yield at large scale in developing countries where crop yield observations in high quality and quantity are challenging to obtain.**

Cornell CALS
College of Agriculture and Life Sciences

Thank You

Overview

- Want to generate tolerably accurate, low cost, predictions of poverty and malnutrition prevalence rates using open source, remotely sensed data and easily implementable methods
- Features:
  - ARENA data (Geographic Features: Location, rurality, ...)
  - Food price data
  - SIF/CHRPS/LST
  - Conflict data

Model Motivation

- Feature set varies across countries
- Outcomes are shared across surveys
- Want to use a simple model which can be fit on individual countries or surveys separately
- Outcomes may be correlated, and we want to investigate / use this

Model Introduction

- Let $x^{t,c}$ denote features drawn from country c, in survey year t
- $y^{t,c}$ denote shared outcomes
- Want to generate estimates for survey specific mappings:

$$y^{t,c} = \hat{f}^{t,c}(x^{t,c}) + \epsilon^{t,c}$$

- $\epsilon^{t,c}$ has mean 0, covariance Σ

MRF Motivation

- Generate function estimates $\hat{f}^{t,c}$ using a Mahalanobis random forest (MRF)
- Motivation for use of MRF:
  - Inexpensive and easy to fit (computationally, practically)
  - Allows for joint estimation of potentially correlated outcomes
  - RFs tend to produce competitive estimates for many problems

Evaluating Model Performance

- Prediction and training done at country specific level
- Two distinct predictive frameworks:
  - Contemporaneous Prediction
  - Forecasting
- Contemporaneous: For each year t, use data from years $t' \leq t$ for training. Data held out from year t used for testing
- Forecasting: For each year t, use only data from years $t' < t$ for training
- Forecasts are generated up to 5 years out

Results: Sequential Forecasting

- Accuracy measured by r^2 and RMSE, normalized by observed outcome range
- Accuracy assessed at three levels of aggregation
  - Fully aggregate across all surveys
  - Country-level
  - Individual survey level

Forecasting: Aggregate Results

Table. Aggregate, out-of-sample r^2 and NRMSE, indexed by methodology and prevalence.

	Child Stunting	Child Wasting	Healthy Weight	Asset Poverty	Underwt Women
IRF r^2	0.07	-0.01	-0.21	0.21	0.31
MRF r^2	0.08	0.10	-0.04	0.21	0.29
IRF NRMSE	0.21	0.15	0.16	0.26	0.17
MRF NRMSE	0.21	0.12	0.15	0.27	0.12

- MRF outperforms independent random forest (IRF) for Wasting, Healthy Weight
- Asset poverty seems much easier to predict than nutrition measures
- NRMSE is quite low for either method, despite weaker r^2

Forecasting: Country level results

Table. Survey size weighted, mean, country level r^2 and NRMSE for forecasting

	Child Stunting	Child Wasting	Healthy Weight	Asset Poverty	Underwt Women
IRF r^2	0.01	-0.24	-0.39	0.12	0.11
MRF r^2	0.02	-0.04	-0.17	0.13	0.10
IRF NRMSE	0.21	0.16	0.17	0.24	0.17
MRF NRMSE	0.21	0.14	0.16	0.24	0.17

- r^2 decreases slightly when assessing predictions at country-level scales (not too serious)
- MRF still shows improvement for wasting, healthy weight
- NRMSE still acceptable for cheap and easy, first pass assessment

Forecasting: Survey Level Results

Table. Survey size weighted survey level r^2 and NRMSE for forecasting

	Child Stunting	Child Wasting	Healthy Weight	Asset Poverty	Underwt Women
IRF r^2	0.00	-0.25	-0.38	-0.29	0.09
MRF r^2	0.01	-0.05	-0.17	-0.12	0.07
IRF NRMSE	0.21	0.15	0.16	0.26	0.17
MRF NRMSE	0.21	0.14	0.15	0.26	0.17

- r^2s are much worse at survey level scales, NRMSE stable
- Likely the result of small sample size in early testing years without data

Forecasting: Summary and Comparison

- r^2 is relatively low for malnutrition indicators, and decays at finer assessment scales
- Related works for forecasting are sparse, closest is Yeh et al. (2020)
  - Aggregate and mean country-level $r^2 \simeq .18$
  - Relative to our $.21, .14$
- MRFs can generate reasonably accurate predictions of povery and malnutrition status, up to 5 years in advance, as measured by NRMSE
- Very easy to implement and use
- Predictions are comprable with neural net approaches, especially at granular assessment scales
- However, sensitivity to sample size is a shortcoming of our method relative to transfer learning approaches

Contemporaneous Prediction

- r^2 is much higher for contemporaneous prediction, similar trends
- NRMSE's qualitatively similar to forecasting
- Joint estimation does not enhance contemporaneous prediction
- Results again weaker at survey level (small sample size)

Contemporaneous Prediction

Table. Comparison to related works for contemporaneous prediction of DHS derived malnutrition and poverty indicators.

Paper	Child Stunting	Child Wasting	Asset Poverty	Underwt Women
Jean *et al.*			0.59, 0.55-0.75†	
Head *et al.*	0.03-0.35†	-0.02-0.11†	0.51-0.74†	0.31-0.47†
Yeh *et al.*			0.67, 0.70*	
Us	0.28, 0.21*, 0.17†	0.23, 0.09*, 0.08†	0.58 , 0.49*, 0.44†	0.48 , 0.24*, 0.20†

Unmarked, $\star$, and † flagged r^2 values represent aggregate, mean country, and mean survey level, results respectively.

- Relative to deep/transfer learning approaches, RFs do about as well in aggregate and at country level

Variable Importance

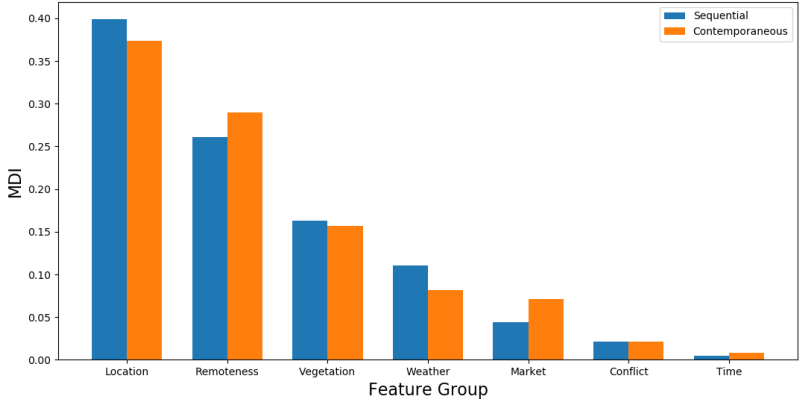
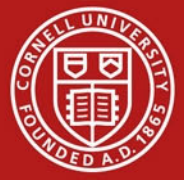


Fig. Feature importance grouped by data type

- Location is key feature for model (somewhat problematic)

Summary

- MRFs can produce inexpensive estimates of poverty and malnutrition status when used in conjunction with our data
  - Estimates are tolerably accurate for first pass systems, particularly at coarse spatial scales where they compete with deep/tfl approaches, and can be produced years in advance
  - Are weaker at fine spatial scales, largely due to sample size issues
- Poverty is easier to predict than malnutrition
- Joint estimation offers some improvement in forecasting
- Location and remoteness data seems most useful, followed but meteorological and vegetative data



Cornell University

Harnessing Big Data and Machine Learning to Monitor Food Security and Malnutrition in Kenya

- Research question: We aim to predict malnutrition and food security outcomes across space and over time at the ward/community level in Kenya.

- Method: Random forest for time-series prediction

- Key innovation: rely on a large-scale, long-term, high-frequency household survey data collected by the National Drought Management Authority (NDMA) to train and validate our machine learning model.

- Potential policy impacts: can identify promising new analyses that NDMA might be able to employ to inform early warning and emergency response in Kenya, especially in its drought-prone regions. Also identify how forecast performance degrades as push further into future.

The NDMA dataset:

- Two sets of panel data: **2006-2016** and **2016-2020**

- Middle Upper Arm Circumference **(MUAC)** as acute malnutrition indicator

- Temporal resolution: **monthly**

- Spatial resolution:

  - Average at the **ward level** - Meaningful administrative level from a targeting standpoint

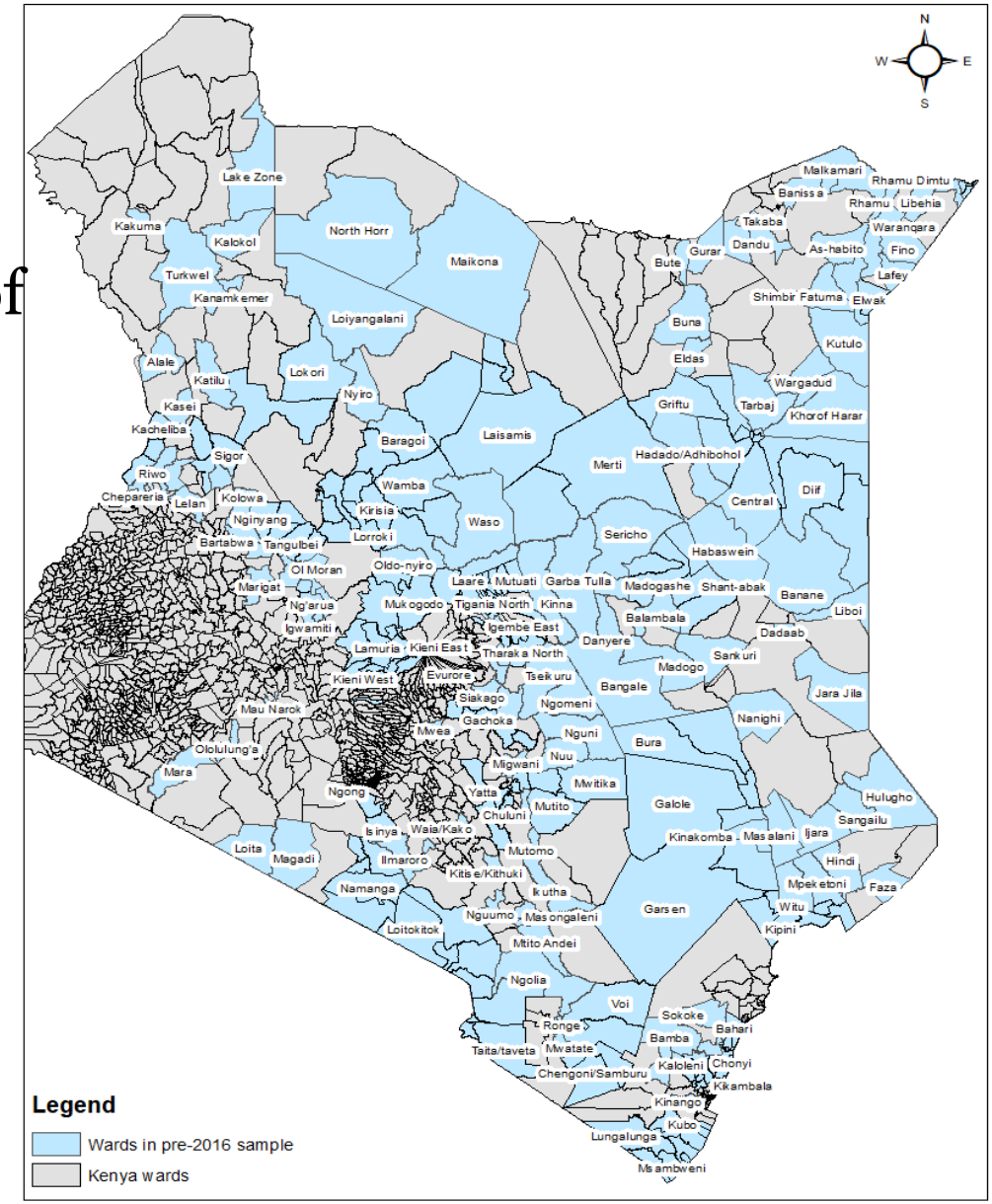
Cornell University

2006 – 2016 dataset

- Data gathered under pre-2010 constitutional reform of administrative levels

Geographic coverage:

- **113** singly identified **wards**
- **66** old **divisions** (comprise multiple current wards)
- Median of 297 monthly observations per ward/division (ranging from 61 to 2,878)

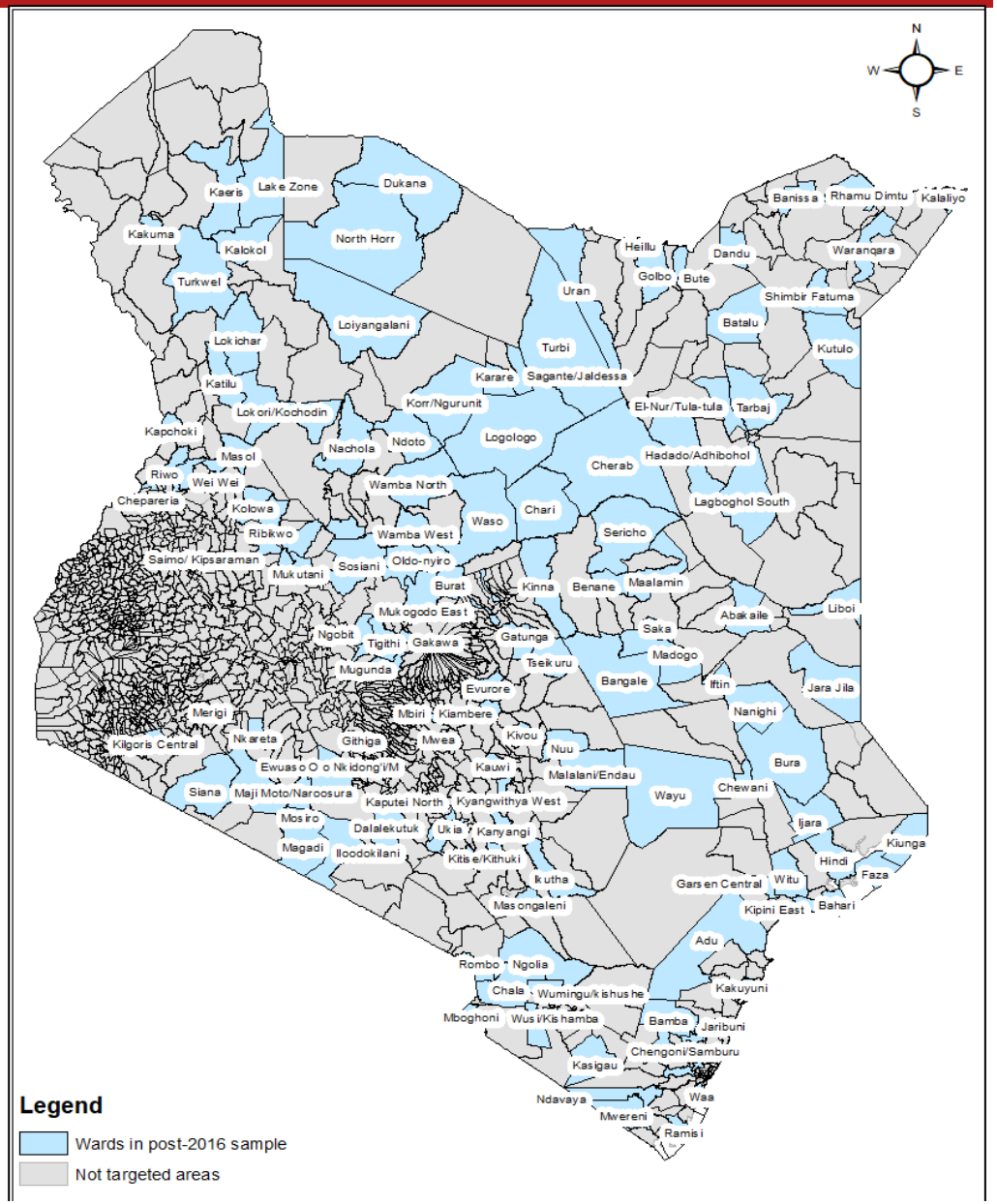


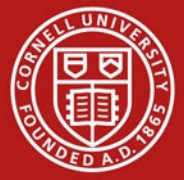
Cornell University

2016 – 2020 dataset

Geographic coverage:

- **157** wards
- Median of 131 monthly observations per ward (ranging from 35 to 357)

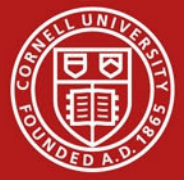




Cornell University

Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning

Linden McBride, Christopher B. Barrett, Christopher Browne, Leiqiu Hu, Yanyan Liu, David Matteson, Ying Sun, and Jiaming Wen



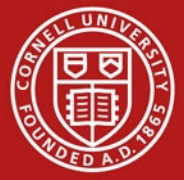
Cornell University

Trade offs, innovations, frontiers

Can big data revolutionize poverty and malnutrition mapping, targeting, M&E and forecasting?

Agencies' needs vary; consider purpose, use case, of the tool

- What type of deprivation is being mapped/targeted/ monitored/forecasted?
- What is the time horizon?
- How transparent/accessible does final model need to be?
- How onerous is the data collection and curation task?
- What is the agency's objective function (Zhou et al. 2021)?



Cornell University

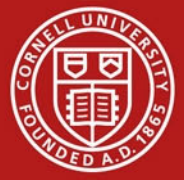
Fit tools to tasks

- **Targeting versus mapping**: Targeting identifies poor/malnourished *people*; mapping identifies *places*

- **Structural versus stochastic**: Tension bn asset-based theory, and empirics, of poverty traps and big data/ML

- **Static versus dynamic**: Mapping/targeting efforts tend to produce static models; early warning systems identify those who will be poor/malnourished/food insecure next period/anticipate shocks' impact on vulnerable pops

- **Data needs**: ML informed map, tool, or model will only be as good as the data on which it is trained and tested

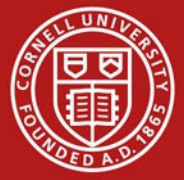


Cornell University

Targeting innovations

Scorecard approach to proxy means test development using machine learning for dimension reduction and out of sample validation for model assessment

- Lean data (Schriener 2007, Kshirsagar et al. 2017, Baez et al. 2019)
- High frequency data (Knippenberg et al. 2019)
- Administrative data (Altındağ et al. 2021)

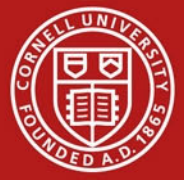


Cornell University

Mapping innovations

Combining different data inputs using CNN or other ML methods to estimate measures of deprivation at local levels

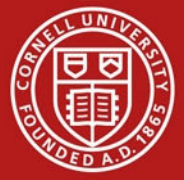
- Cell Data Records (Blumenstock et al. 2015)
- Nightlights, daytime satellite imagery, NDVI, remotely sensed data (Jean et al. 2016, Yeh et al. 2020)
- Combined data sources (Pokhriyal & Jacques 2017, Yeh et al. 2020)
- Multidimensional Poverty Index (Pokhriyal & Jacques 2017, Njuguna & McSharry 2017)
- Open source data (Hersh et al. 2020)
- Multivariate prediction of correlated outcomes (Browne et al. 2021)



Cornell University

Targeting and mapping frontiers

- Multi-dimensional nature of deprivation
- Predicting low probability/noisy outcomes (Head et al. 2017)
- Limitations of available data (Blumenstock 2016 & 2020)
- Determinants of geographically concentrated poverty (Yeh et al. 2020)



Cornell University

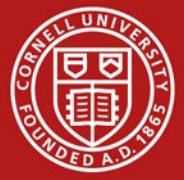
Structural versus stochastic

Innovations

- Recent mapping models do relatively well predicting asset poverty across space (Blumenstock et al. 2015, Jean et al. 2016, Yeh et al. 2020, Browne et al. 2021) and over time (Yeh et al. 2020, Browne et al. 2021)

Frontiers

- Parsing persistently poor from the dynamically mobile (Carter & Barrett 2006)
- Better mapping well-being dynamics
- Resilience targeting/ resilience mapping (Barrett and Constas 2014, Cisse & Barrett 2018, Upton, Cisse & Barrett 2016, Knippenberg et al. 2019)



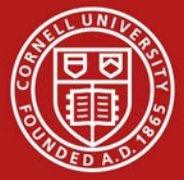
Cornell University

Static versus dynamic innovations

- Food insecurity early warning using high frequency data (Mude et al. 2009, Lentz et al. 2018)

- Tang et al. (2018), Yeh et al. (2020) demonstrate that CNNs trained on changes in satellite imagery can predict changes in consumption or asset wealth in future periods

- Browne et al. (2021) produce contemporaneous and sequential prediction of correlated asset wealth and malnutrition indicators



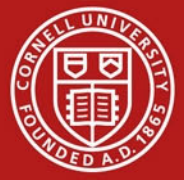
Cornell University

Static versus dynamic frontiers

- High frequency data needed for monitoring and early warning (Barrett 2010, Headey & Barrett 2015)

- Focus on prediction of changes (not just levels)

- Integration of time series statistics with ML tools with application in these settings



Cornell University

Data

- Undersupply of the global public good of collection, standardization, updating and open access curation of key variables

- One can only reliably predict states and processes that have been previously observed in data → assumed stationarity in DGP

- COVID has likely accelerated trends towards more creative data collection (Blumenstock 2020)

Summary

- Big data and ML methods are revolutionizing mapping, targeting, M&E, and early warning
- However, effective use requires thoughtful consideration of the purpose and use cases of the map/tool/model
- Data availability and curation remain a serious limitation

Cornell University

Thank you

Thank you for your interest
Comments/questions?