Landsat Image Time Series Processing using HTCondor on UW-CHTC and OSG Resources

Matthew Garcia, Ph.D.
Prof. Philip A. Townsend

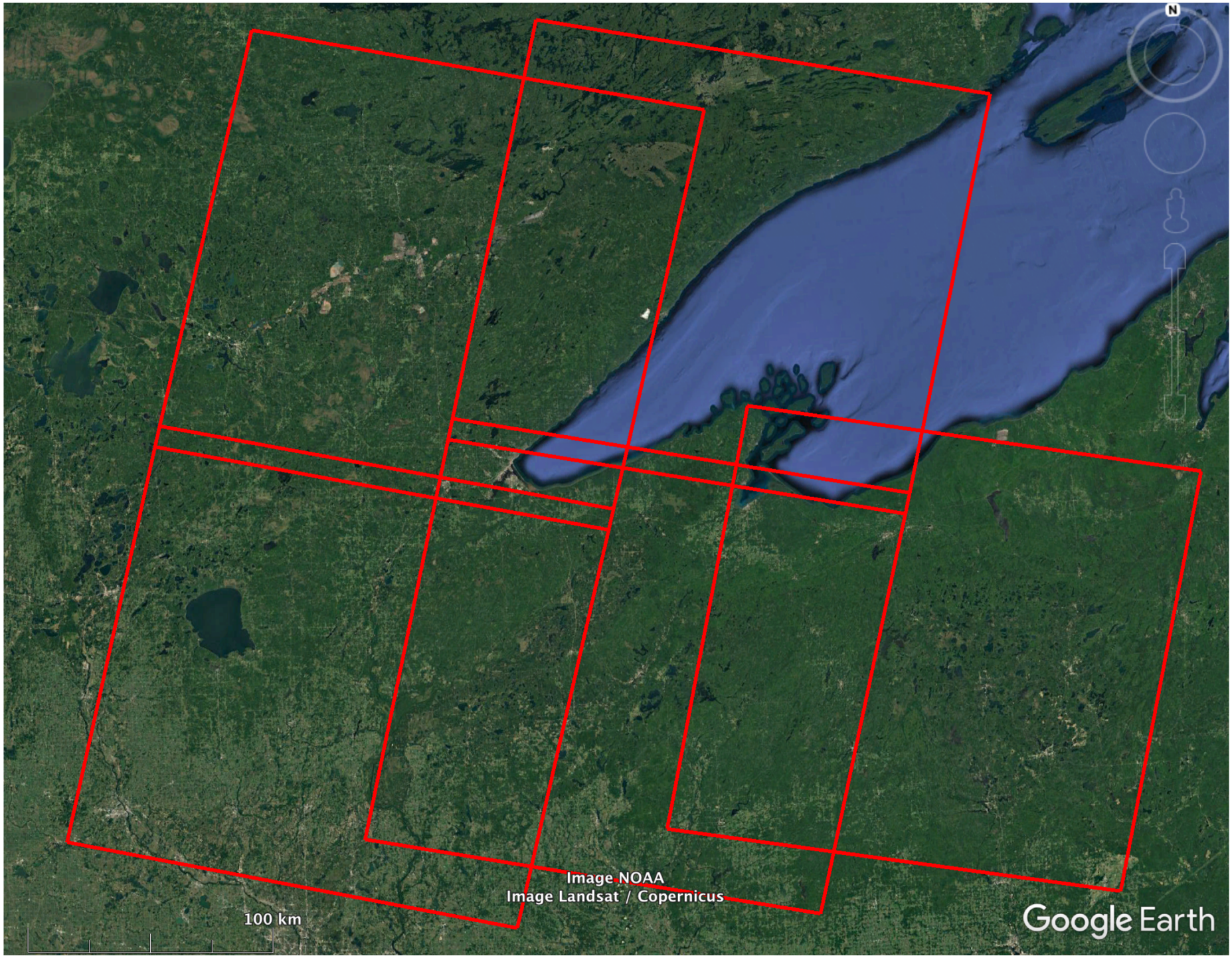
Dept. of Forest & Wildlife Ecology
University of Wisconsin–Madison

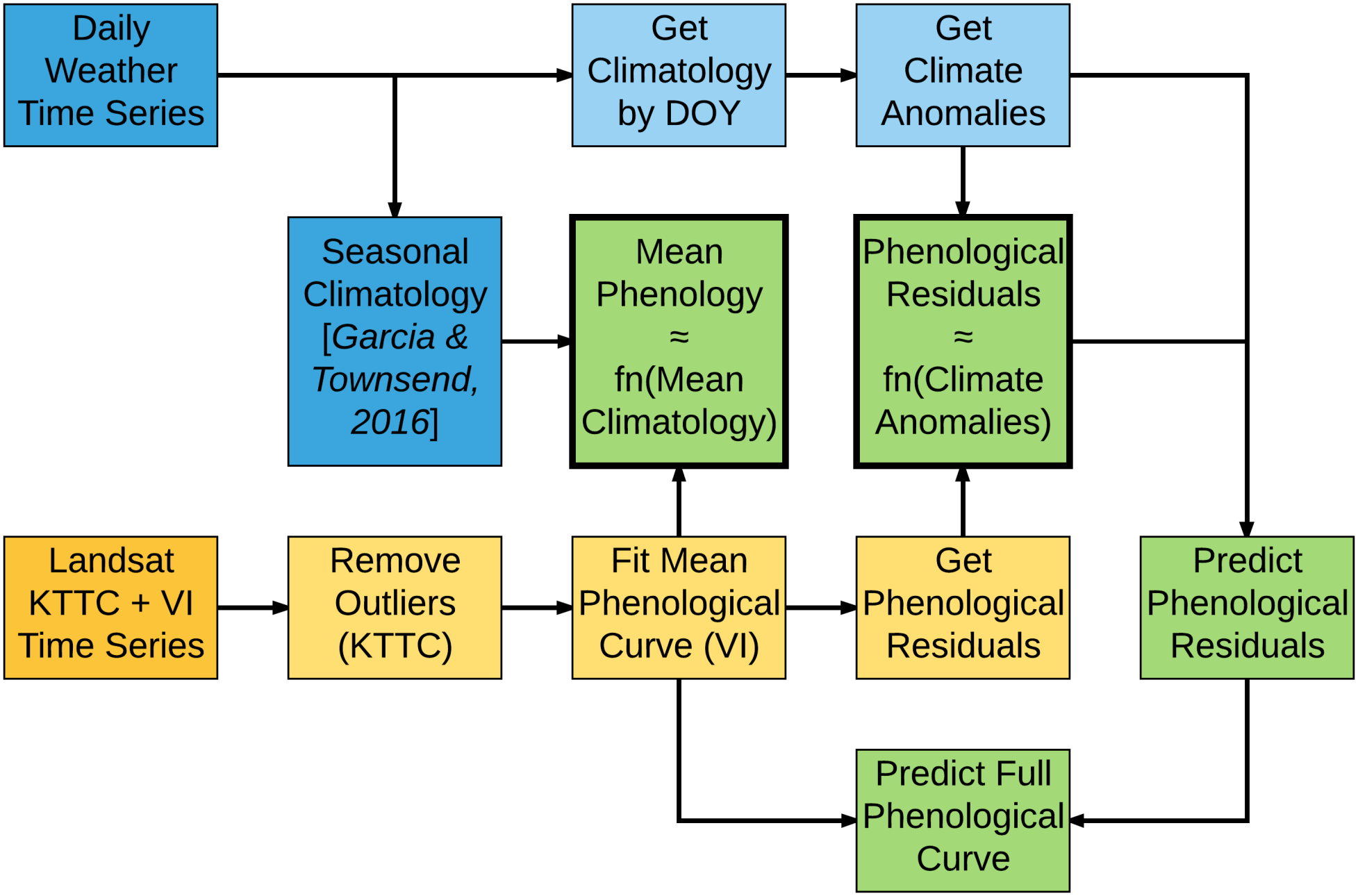
HTCondor Week
24 May 2018

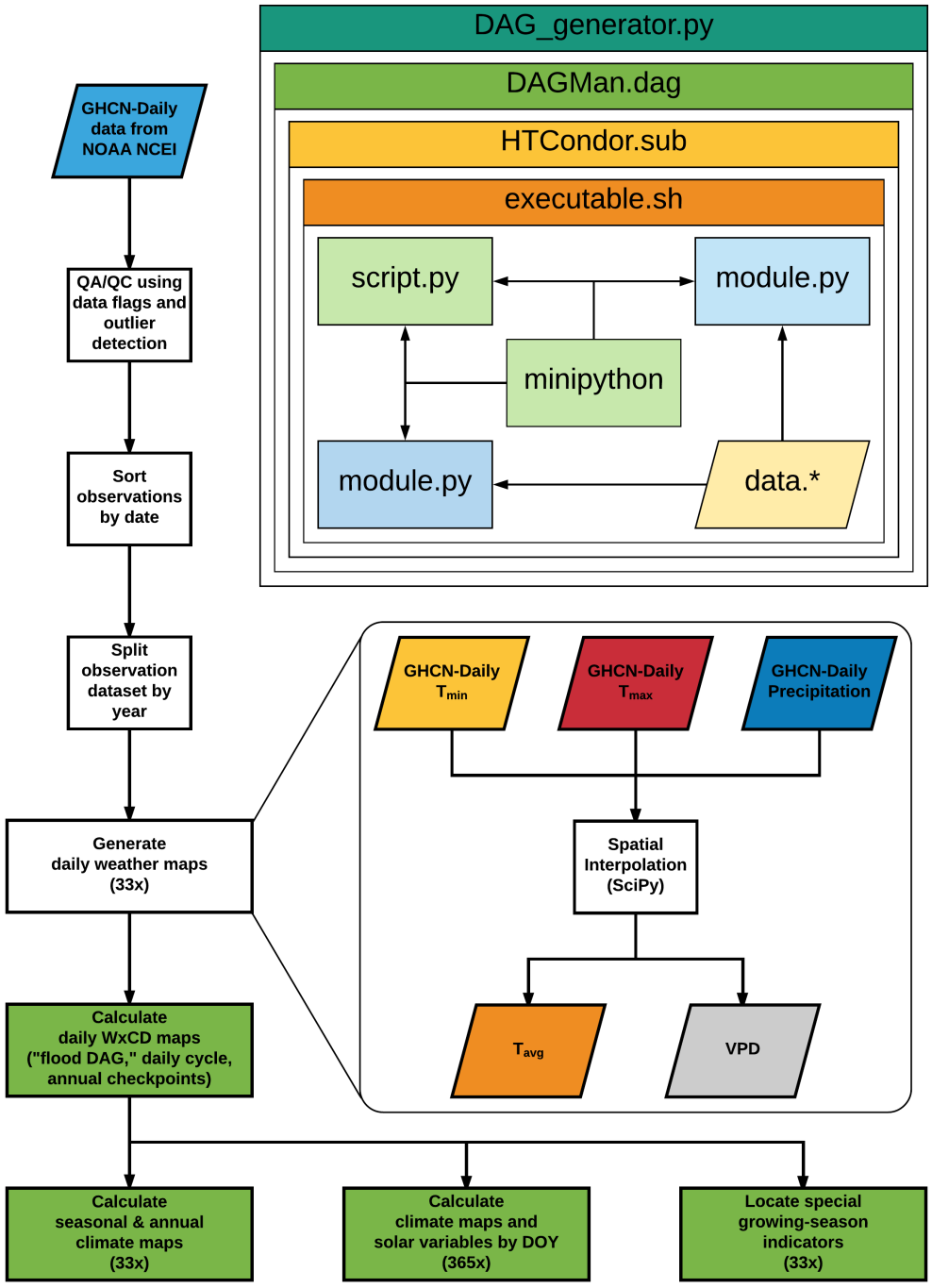
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

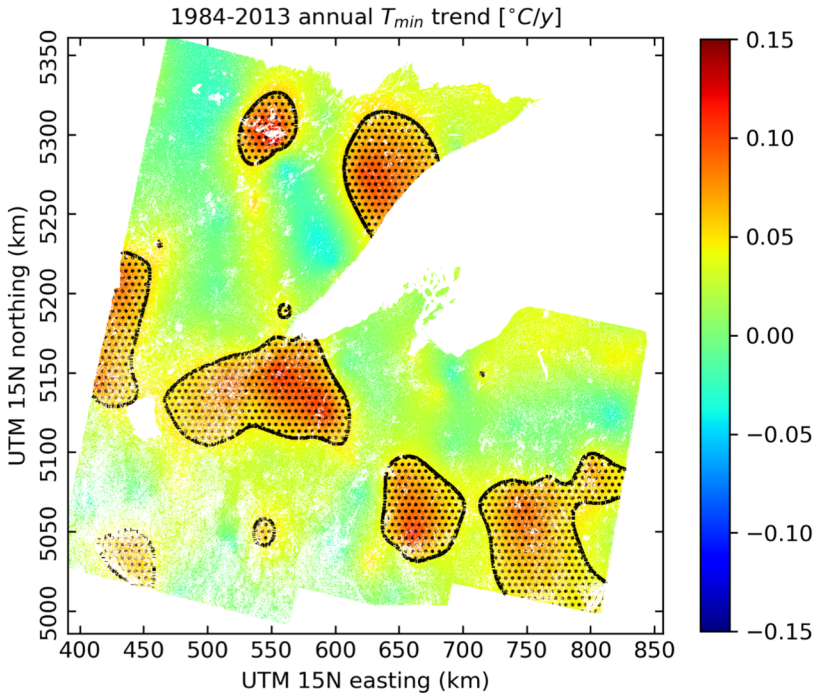
CENTER FOR
HIGH THROUGHPUT
COMPUTING

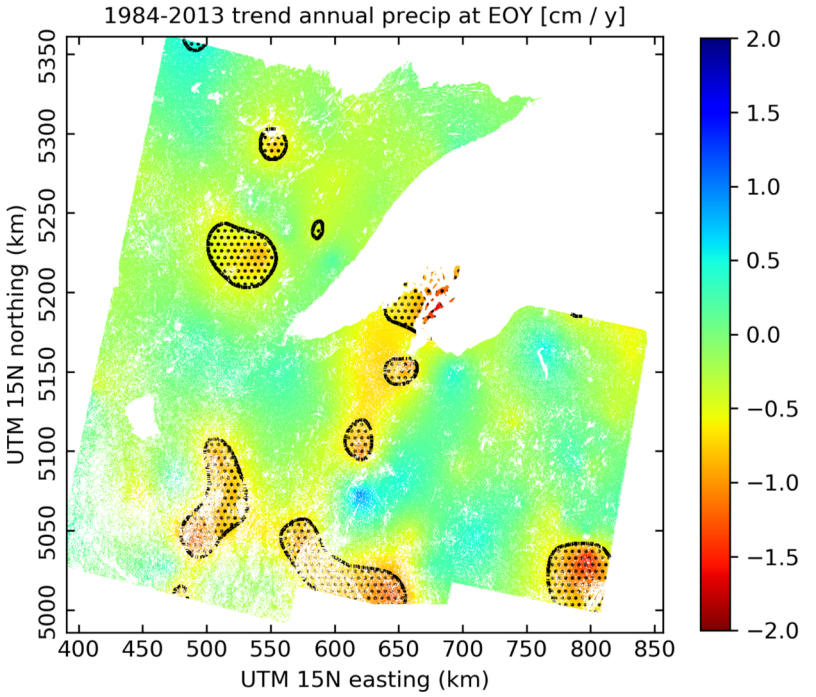
Open Science Grid


N
Image NOAA
Image Landsat / Copernicus
100 km
Google Earth

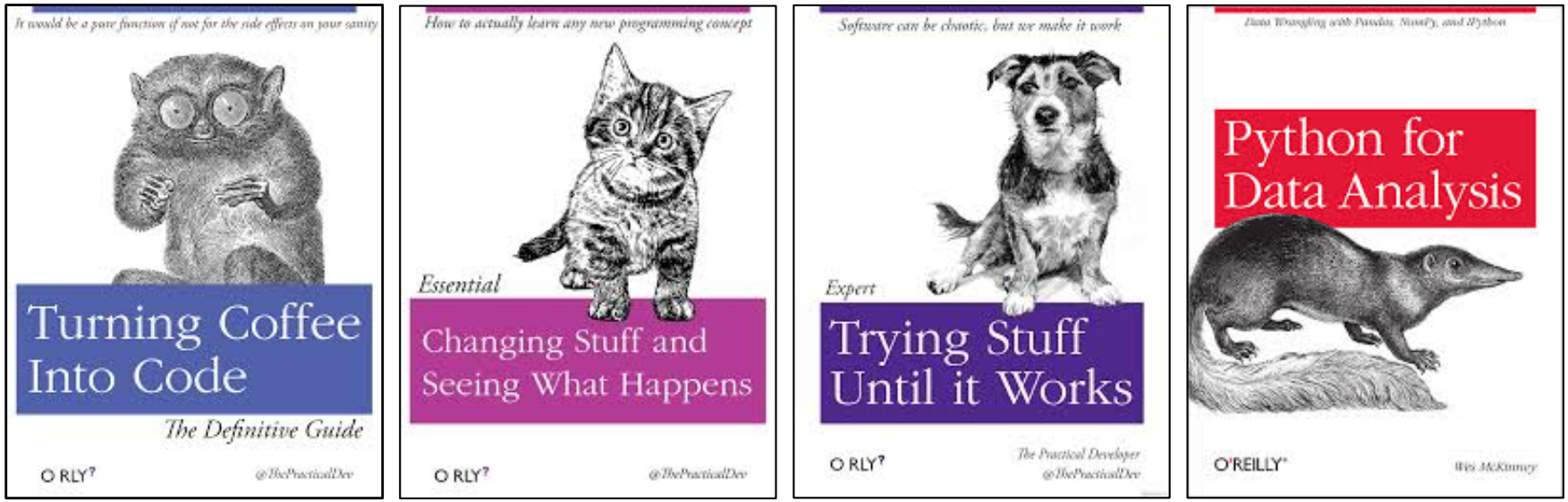

Daily Weather Time Series
Get Climatology by DOY
Get Climate Anomalies
Seasonal Climatology [*Garcia & Townsend, 2016*]
Mean Phenology ≈ fn(Mean Climatology)
Phenological Residuals ≈ fn(Climate Anomalies)
Landsat KTTC + VI Time Series
Remove Outliers (KTTC)
Fit Mean Phenological Curve (VI)
Get Phenological Residuals
Predict Phenological Residuals
Predict Full Phenological Curve

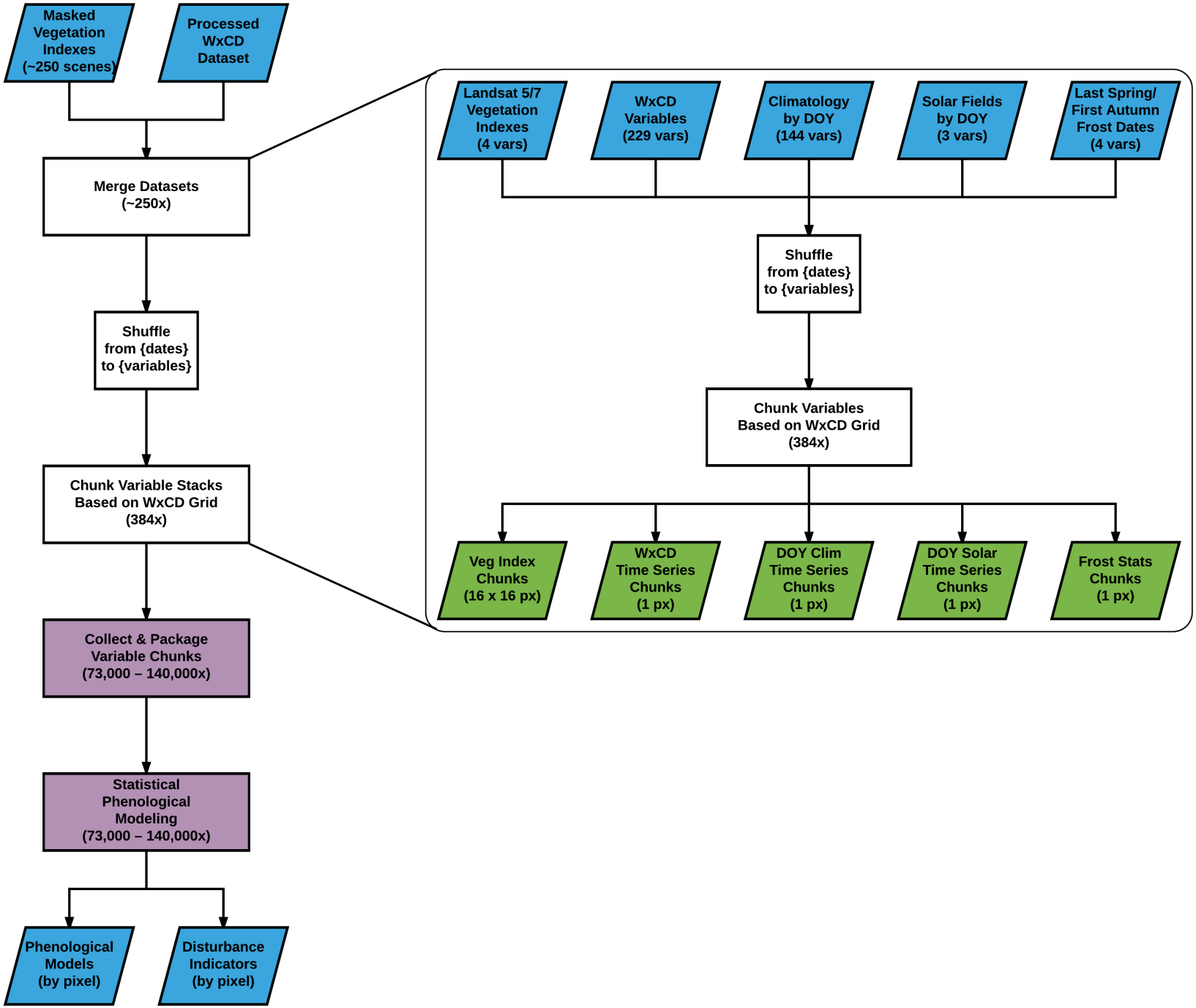

DAG_generator.py
DAGMan.dag
HTCondor.sub
executable.sh
script.py
module.py
minipython
module.py
data.*
GHCN-Daily data from NOAA NCEI
QA/QC using data flags and outlier detection
Sort observations by date
Split observation dataset by year
Generate daily weather maps (33x)
GHCN-Daily T_{min}
GHCN-Daily T_{max}
GHCN-Daily Precipitation
Spatial Interpolation (SciPy)
T_{avg}
VPD
Calculate daily WxCD maps ("flood DAG," daily cycle, annual checkpoints)
Calculate seasonal & annual climate maps (33x)
Calculate climate maps and solar variables by DOY (365x)
Locate special growing-season indicators (33x)

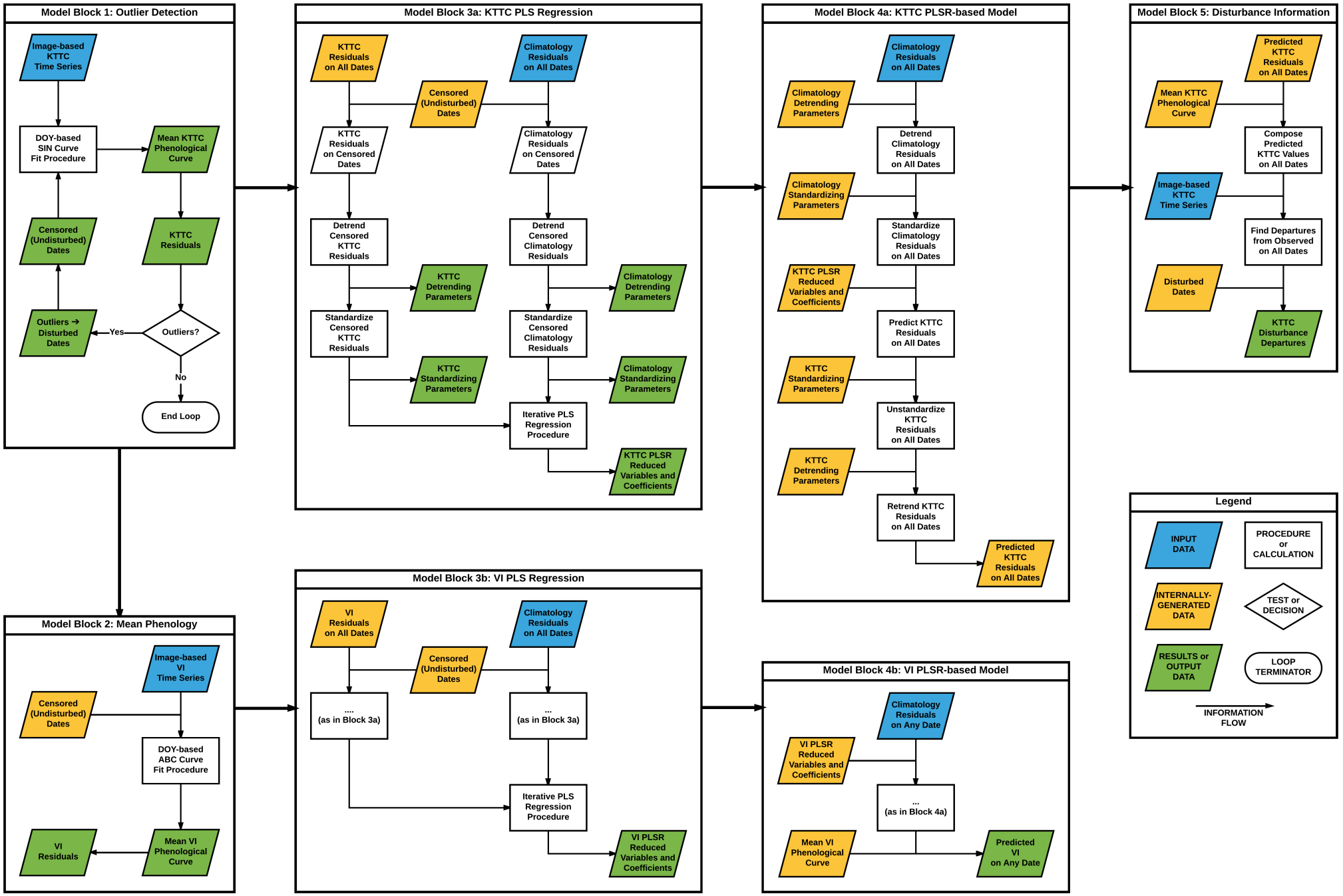

1984-2013 annual T_{min} trend [°C/y]
UTM 15N northing (km)
UTM 15N easting (km)
5000
5050
5100
5150
5200
5250
5300
5350
400
450
500
550
600
650
700
750
800
850
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15

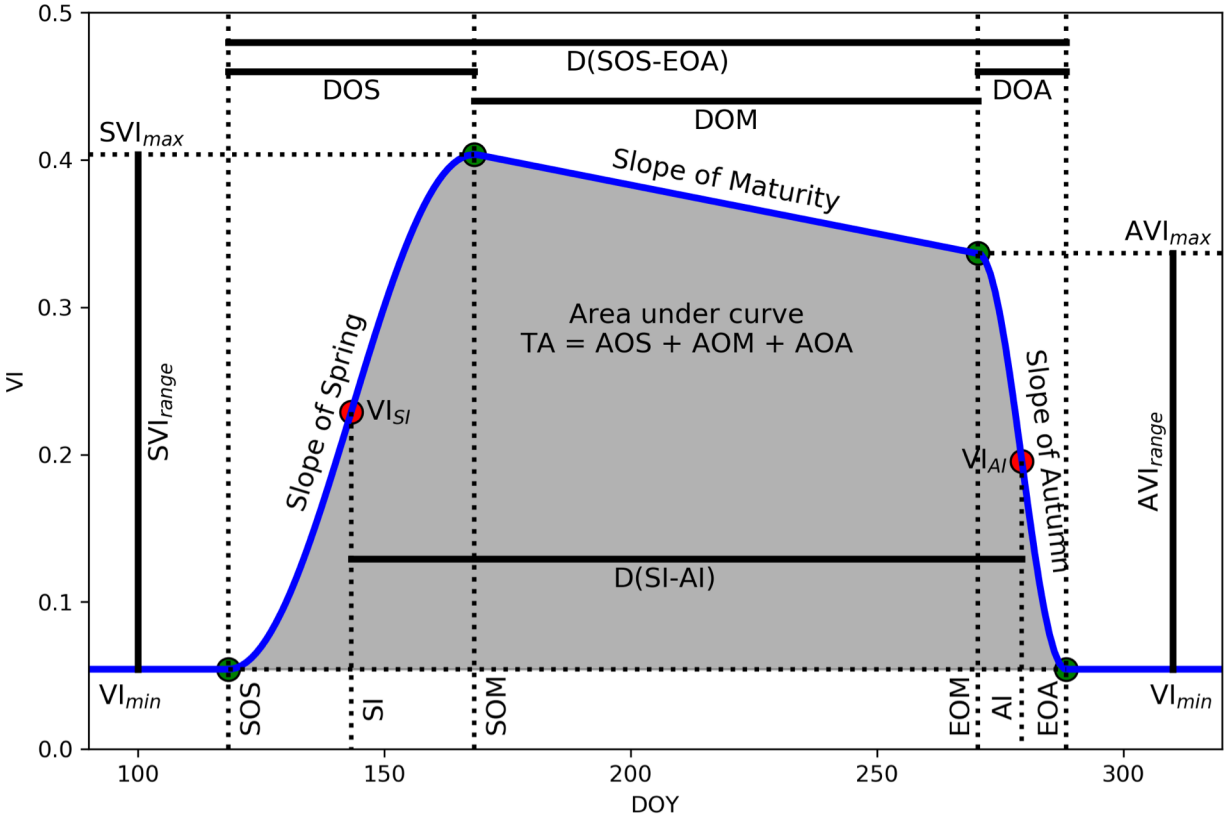

1984-2013 trend annual precip at EOY [cm / y]
UTM 15N northing (km)
UTM 15N easting (km)
5000
5050
5100
5150
5200
5250
5300
5350
400
450
500
550
600
650
700
750
800
850
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0

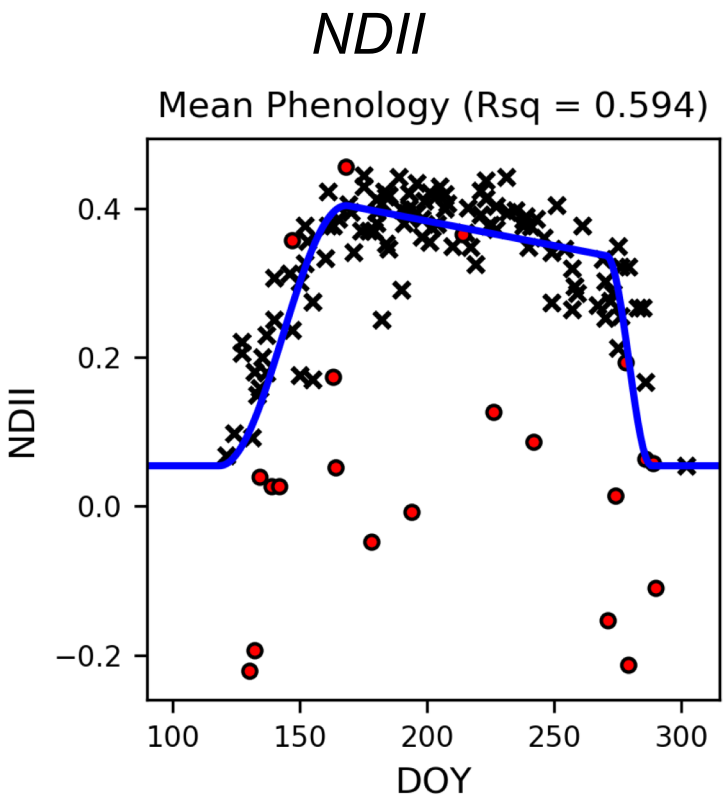
So you think you need to model your data…




Masked Vegetation Indexes (~250 scenes)
Processed WxCD Dataset
Merge Datasets (~250x)
Shuffle from {dates} to {variables}
Chunk Variable Stacks Based on WxCD Grid (384x)
Collect & Package Variable Chunks (73,000 – 140,000x)
Statistical Phenological Modeling (73,000 – 140,000x)
Phenological Models (by pixel)
Disturbance Indicators (by pixel)
Landsat 5/7 Vegetation Indexes (4 vars)
WxCD Variables (229 vars)
Climatology by DOY (144 vars)
Solar Fields by DOY (3 vars)
Last Spring/ First Autumn Frost Dates (4 vars)
Shuffle from {dates} to {variables}
Chunk Variables Based on WxCD Grid (384x)
Veg Index Chunks (16 x 16 px)
WxCD Time Series Chunks (1 px)
DOY Clim Time Series Chunks (1 px)
DOY Solar Time Series Chunks (1 px)
Frost Stats Chunks (1 px)


Model Block 1: Outlier Detection
Image-based KTTC Time Series
DOY-based SIN Curve Fit Procedure
Mean KTTC Phenological Curve
Censored (Undisturbed) Dates
KTTC Residuals
Outliers → Disturbed Dates
Yes
Outliers?
No
End Loop
Model Block 2: Mean Phenology
Image-based VI Time Series
Censored (Undisturbed) Dates
DOY-based ABC Curve Fit Procedure
VI Residuals
Mean VI Phenological Curve
Model Block 3a: KTTC PLS Regression
KTTC Residuals on All Dates
Climatology Residuals on All Dates
Censored (Undisturbed) Dates
KTTC Residuals on Censored Dates
Climatology Residuals on Censored Dates
Detrend Censored KTTC Residuals
Detrend Censored Climatology Residuals
KTTC Detrending Parameters
Climatology Detrending Parameters
Standardize Censored KTTC Residuals
Standardize Censored Climatology Residuals
KTTC Standardizing Parameters
Climatology Standardizing Parameters
Iterative PLS Regression Procedure
KTTC PLSR Reduced Variables and Coefficients
Model Block 3b: VI PLS Regression
VI Residuals on All Dates
Climatology Residuals on All Dates
Censored (Undisturbed) Dates
.... (as in Block 3a)
... (as in Block 3a)
Iterative PLS Regression Procedure
VI PLSR Reduced Variables and Coefficients
Model Block 4a: KTTC PLSR-based Model
Climatology Residuals on All Dates
Climatology Detrending Parameters
Detrend Climatology Residuals on All Dates
Climatology Standardizing Parameters
Standardize Climatology Residuals on All Dates
KTTC PLSR Reduced Variables and Coefficients
Predict KTTC Residuals on All Dates
KTTC Standardizing Parameters
Unstandardize KTTC Residuals on All Dates
KTTC Detrending Parameters
Retrend KTTC Residuals on All Dates
Predicted KTTC Residuals on All Dates
Model Block 4b: VI PLSR-based Model
Climatology Residuals on Any Date
VI PLSR Reduced Variables and Coefficients
... (as in Block 4a)
Mean VI Phenological Curve
Predicted VI on Any Date
Model Block 5: Disturbance Information
Predicted KTTC Residuals on All Dates
Mean KTTC Phenological Curve
Compose Predicted KTTC Values on All Dates
Image-based KTTC Time Series
Find Departures from Observed on All Dates
Disturbed Dates
KTTC Disturbance Departures
Legend
INPUT DATA
PROCEDURE or CALCULATION
INTERNALLY-GENERATED DATA
TEST or DECISION
RESULTS or OUTPUT DATA
LOOP TERMINATOR
INFORMATION FLOW



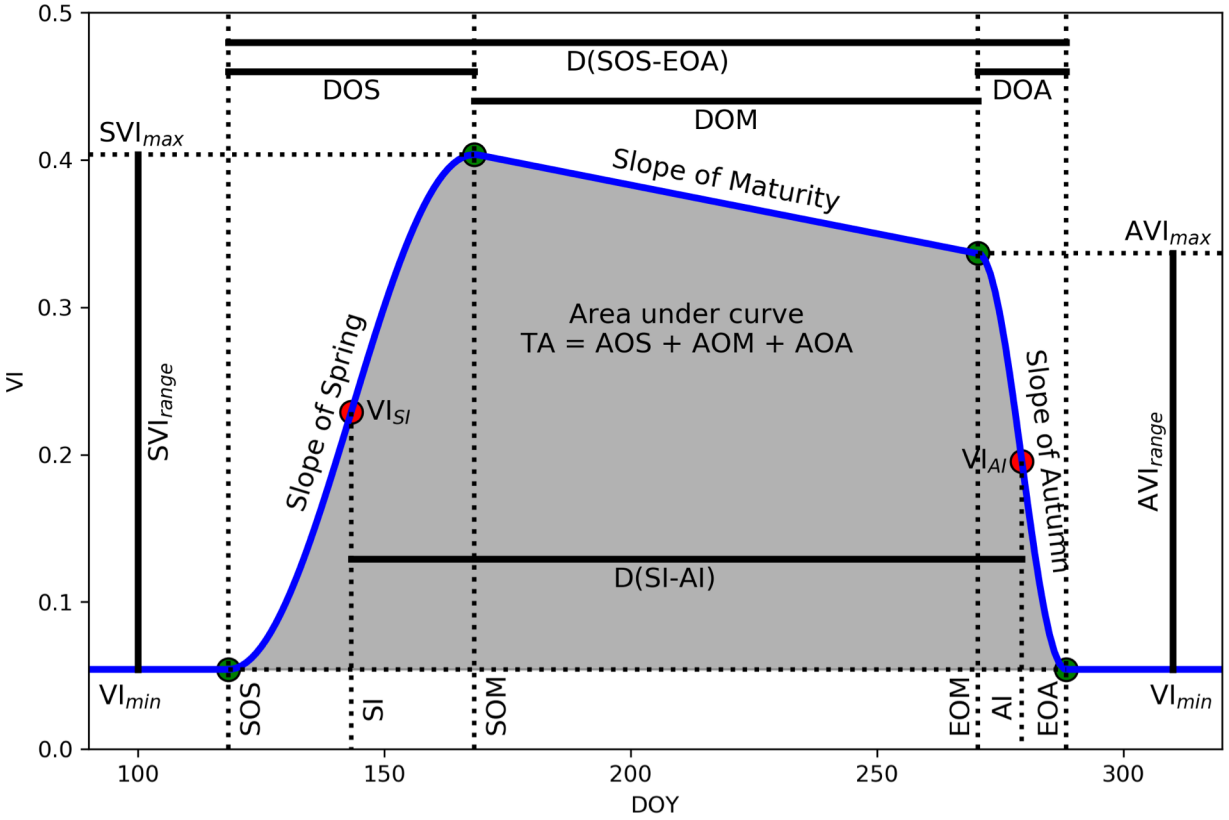


Single pixel time series

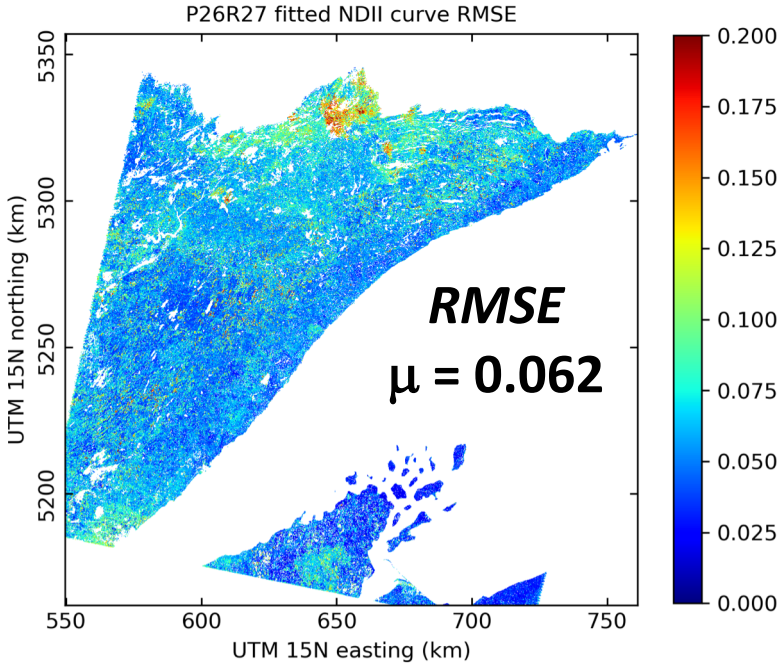
KTTC statistical outliers: red

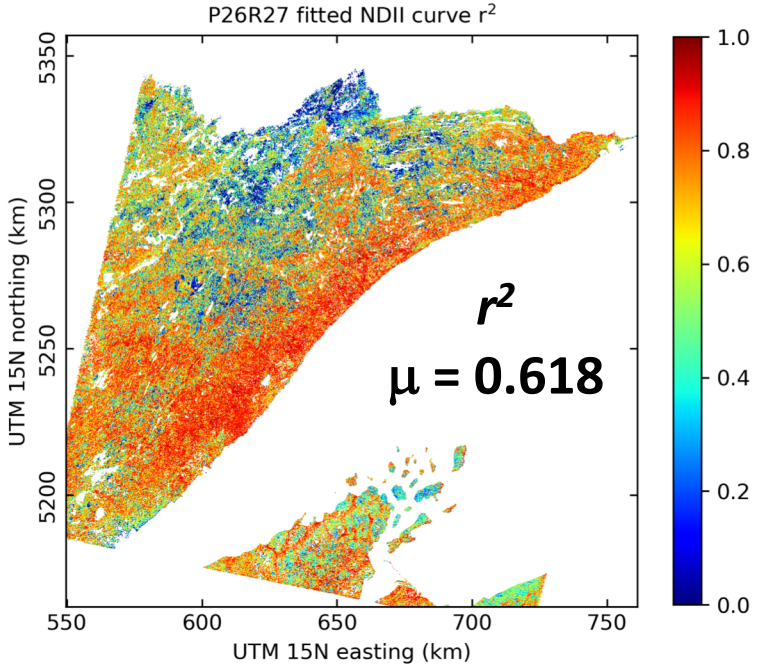
Retained Landsat dates: black

Fitted curve: blue → $r^2 \sim 0.6$



NDII mean phenology





PLS: Projection to Latent Structures, a.k.a.
PLSR: Partial-Least-Squares Regression

Similar to PCA, but…

- maximizes covariance, instead of minimizing correlation
- incorporates the response variable, not just the predictors

Unlike OLS regression, does not assume predictors are error-free
Similar to Multiple Linear Regression, but handles predictor collinearity
→ able to handle many predictor variables with few response variables

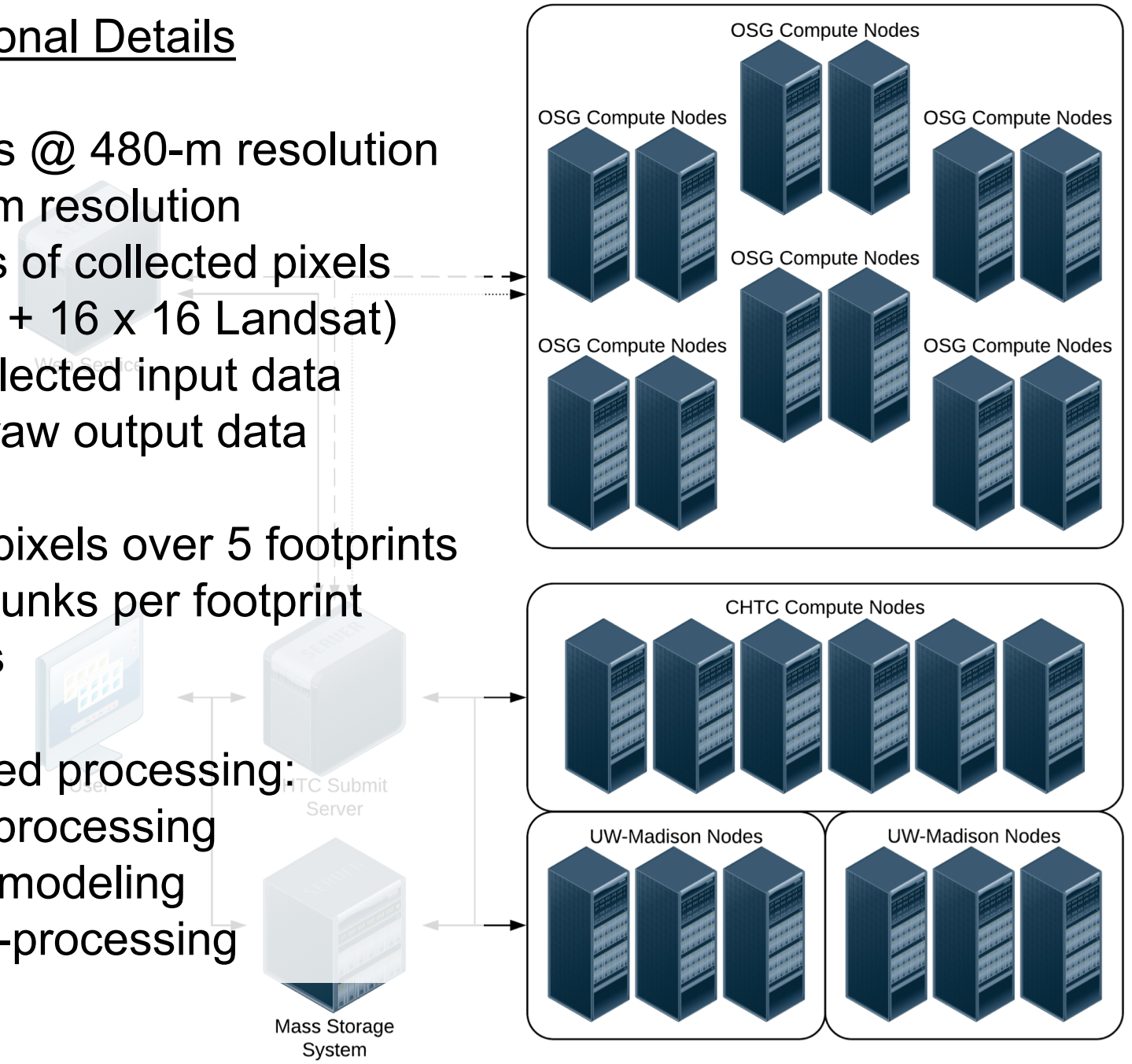
$$\begin{bmatrix} W_{1,1} & \cdots & W_{m,1} \\ \vdots & \ddots & \vdots \\ W_{1,n} & \cdots & W_{m,n} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = \begin{bmatrix} V_1 \\ \vdots \\ V_n \end{bmatrix}$$

Computational Details

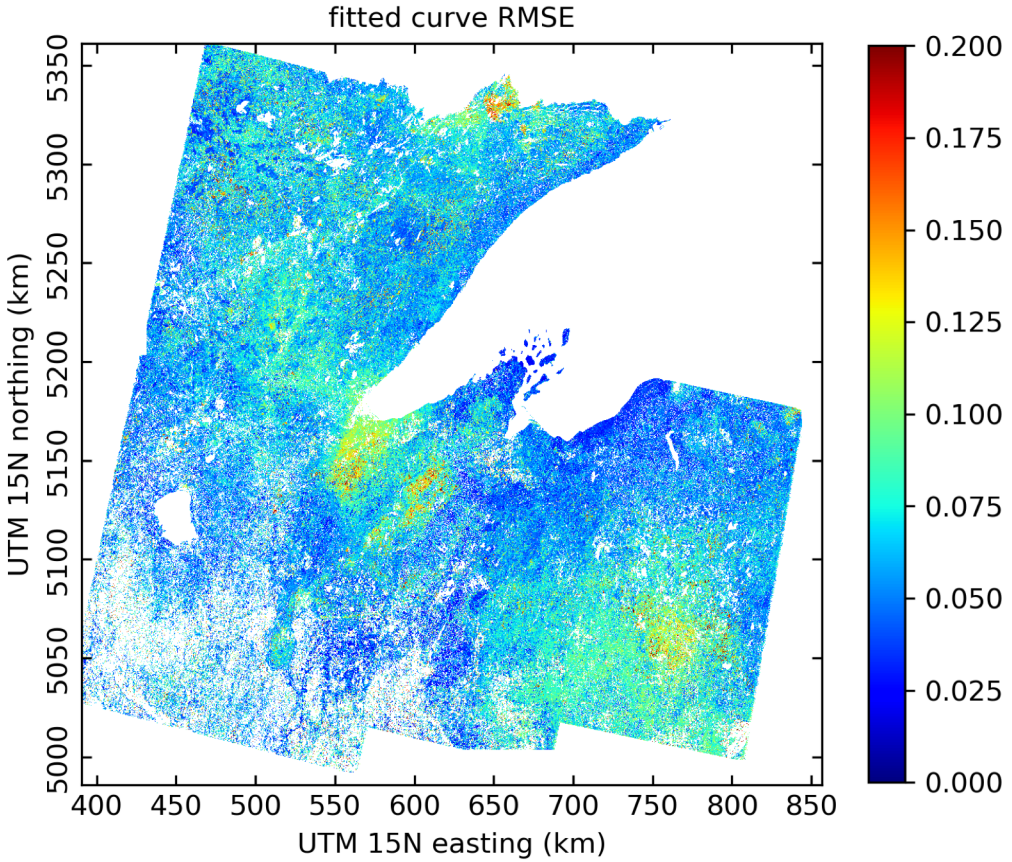
Weather/climate pixels @ 480-m resolution
Landsat pixels @ 30-m resolution
→ Geographic chunks of collected pixels
(1 weather/climate + 16 x 16 Landsat)
~1.5 MB/chunk collected input data
→ ~50 MB/chunk raw output data

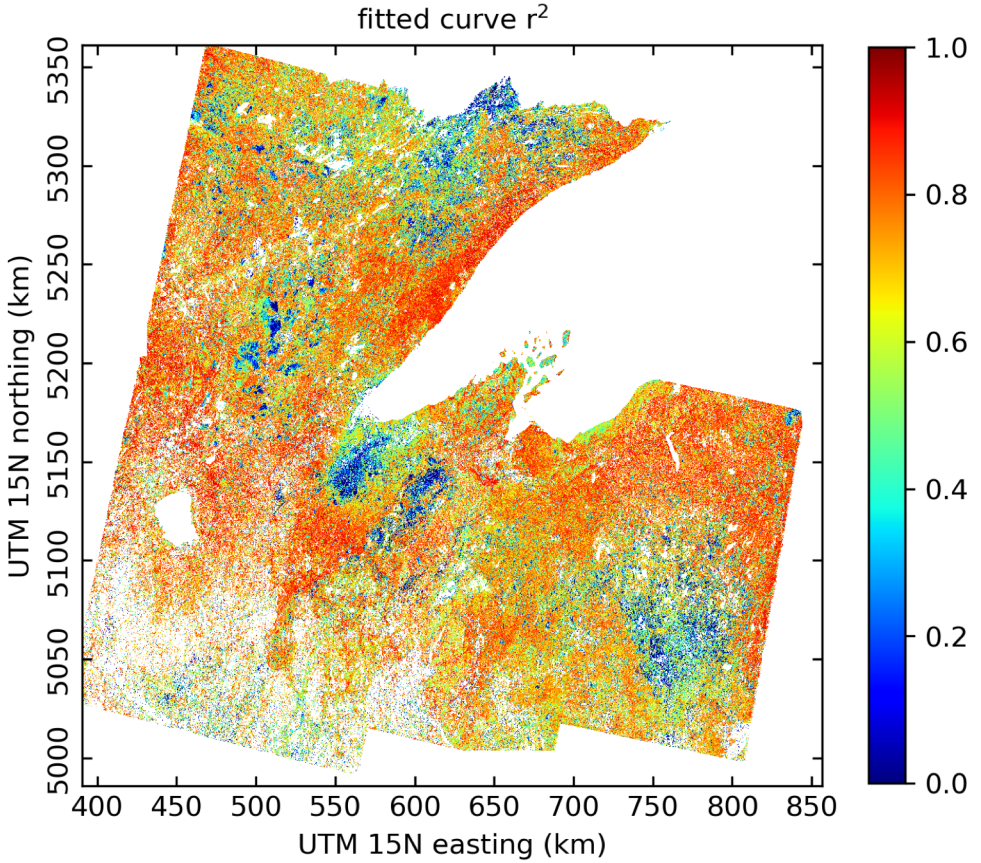
~130 million Landsat pixels over 5 footprints
~70,000 – 140,000 chunks per footprint
~624,000 total chunks

Ideal task for distributed processing:
→ UW CHTC for pre-processing
→ OSG for statistical modeling
→ UW CHTC for post-processing

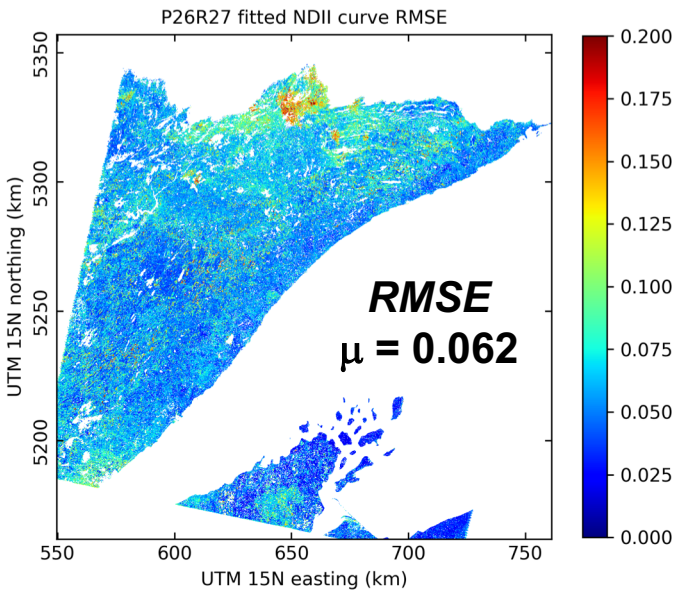


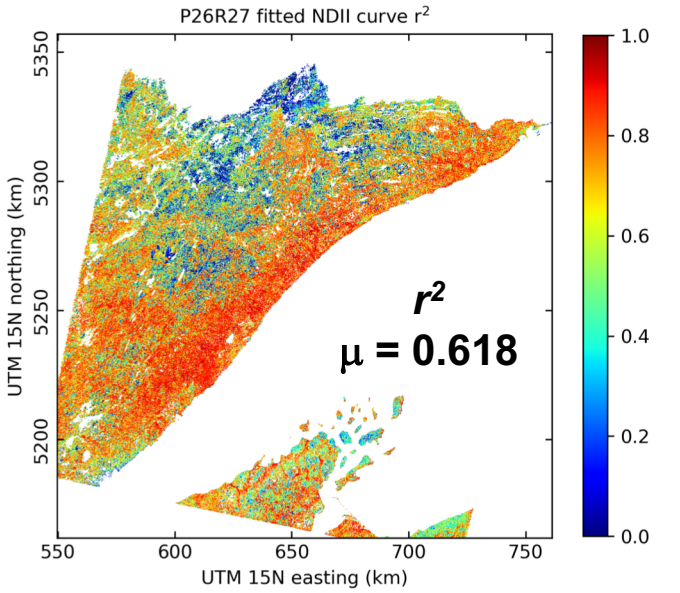
Mean Phenology: *NDII* fitted curve error statistics and goodness-of-fit



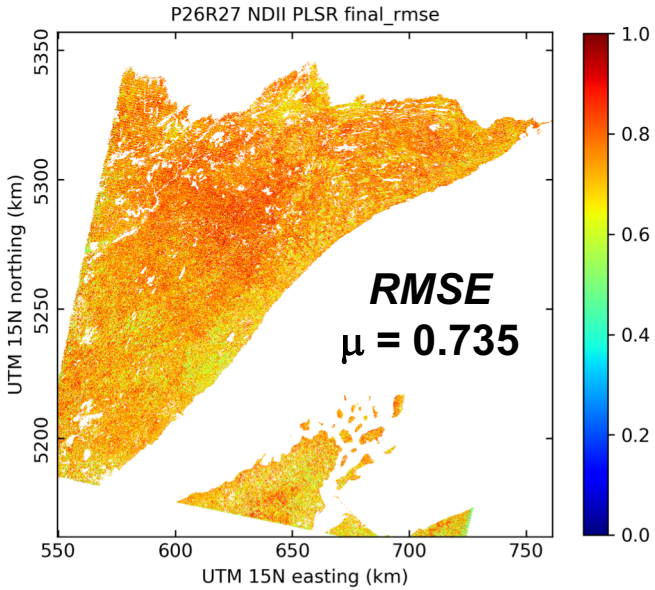


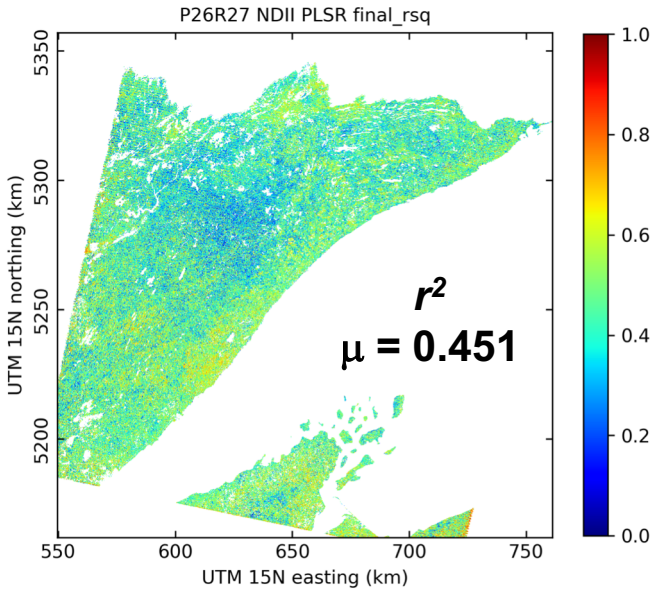
mean phenology



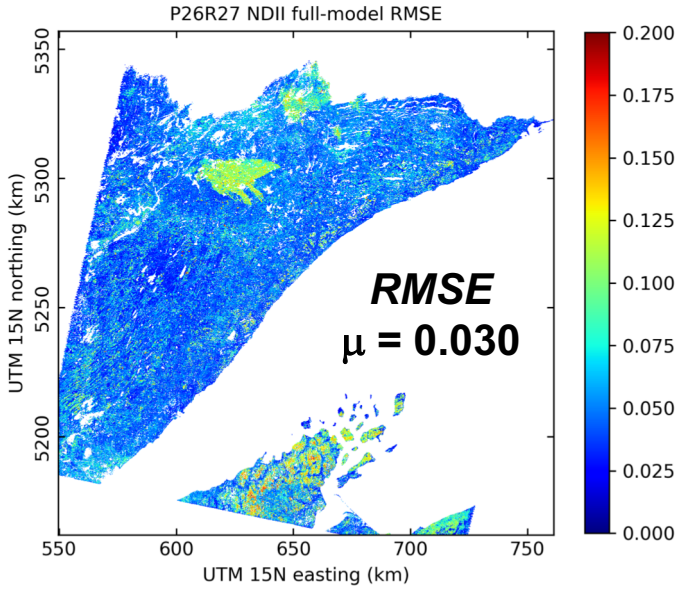


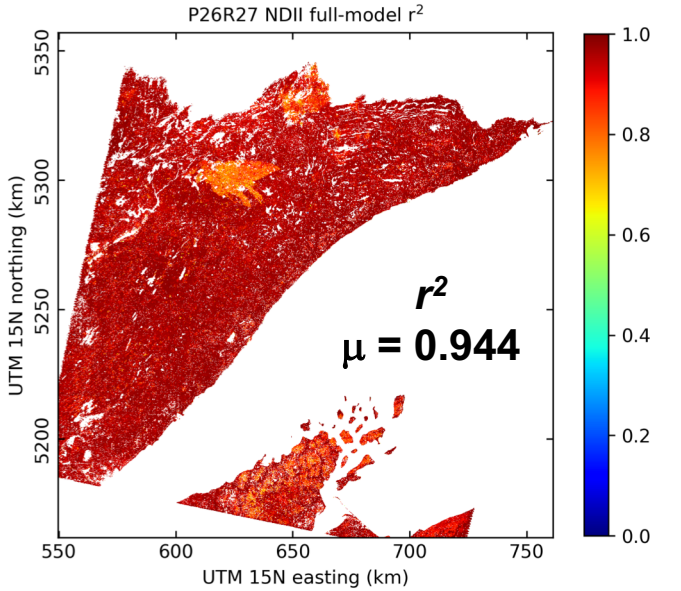
residuals via PLSR





full phenology model





Statistical model:

~624K chunks @ ~12.6 h/chunk = ~8 Mh

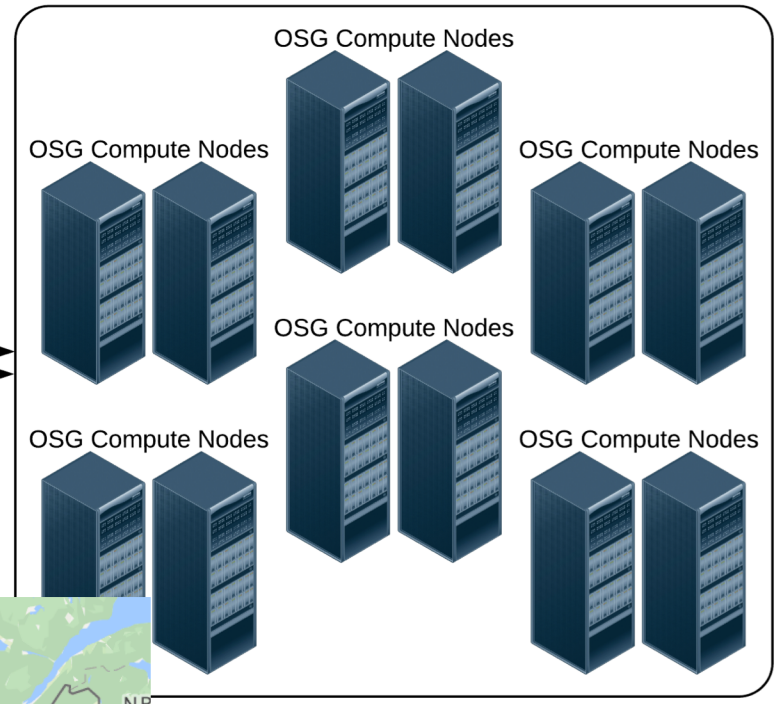
Overall processing time:

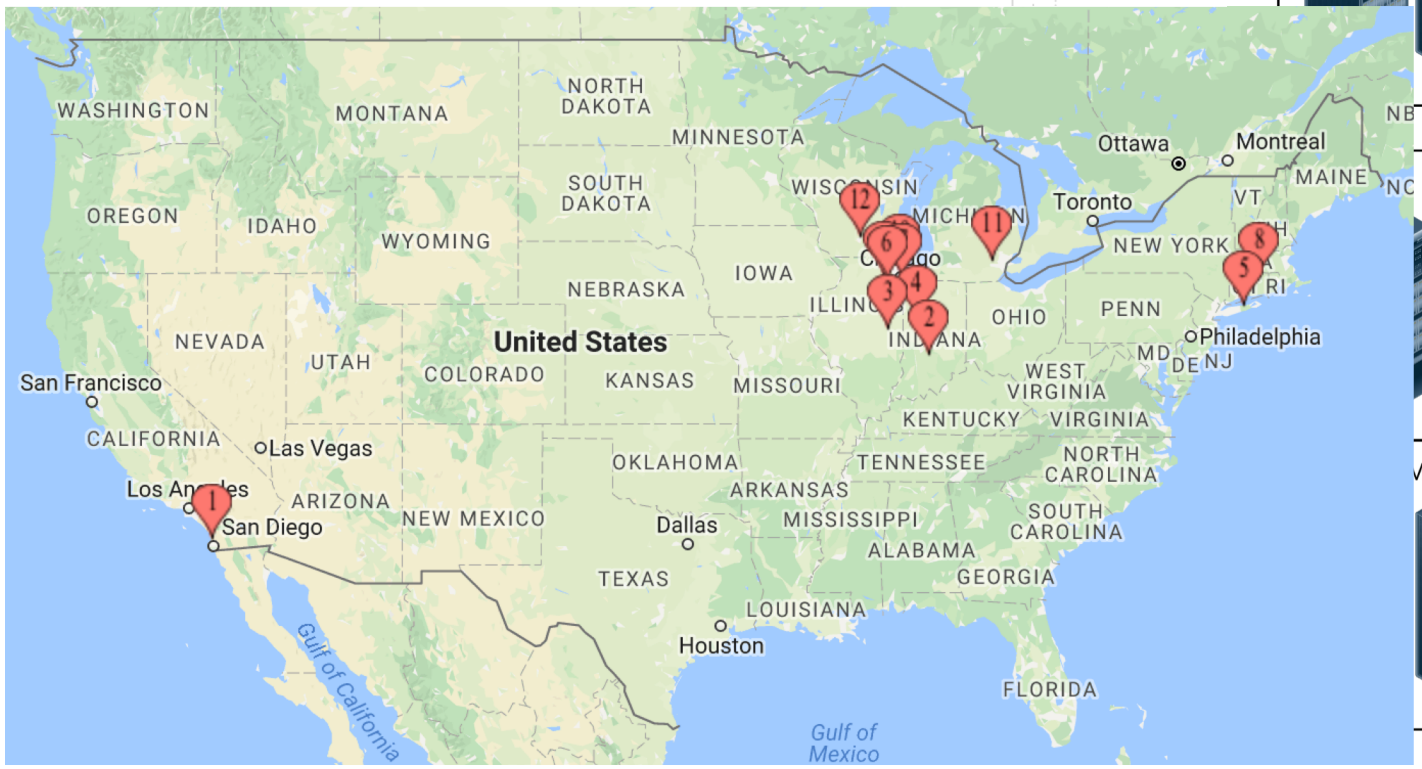
~13 million computing hours

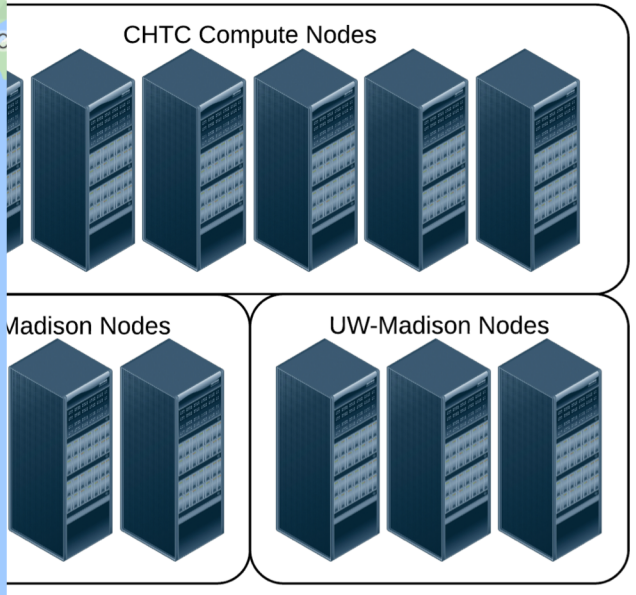
~5.6 Mh on OSG nodes

~5.1 Mh on CHTC resources

~2.3 Mh on other UW clusters







Thank you!

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

HT CENTER FOR HIGH THROUGHPUT COMPUTING

Open Science Grid

Canada
Natural Resources Canada
Canadian Forest Service

http://matthewgarcia.tech