Land surface Verification Toolkit (LVT): A formal benchmarking and evaluation framework for land surface models

Sujay V. Kumar

Hydrological Sciences Laboratory
NASA Goddard Space Flight Center, Greenbelt, MD

Motivation

Entekhabi et al., *BAMS* (1999)

An Agenda for Land Surface Hydrology Research and a Call for the Second International Hydrological Decade

Dara Entekhabi, Ghassem R. Asrar, Alan K. Betts, Keith J. Beven, Rafael L. Bras, Christopher J. Duffy, Thomas Dunne, Randal D. Koster, Dennis P. Lettenmaier, Dennis B. McLaughlin, William J. Shuttleworth, Martinus T. van Genuchten, Ming-Ying Wei, and Eric F. Wood

ABSTRACT

Hydrologic research at the interface between the atmosphere and land surface is undergoing a dramatic change in focus, driven by new societal priorities, emerging technologies, and better understanding of the earth system. In this

Need formal evaluation procedures to improve the "observability" of LSM processes

van den Hurk et al., *BAMS* (2011)

ACCELERATION OF LAND SURFACE MODEL DEVELOPMENT OVER A DECADE OF GLASS

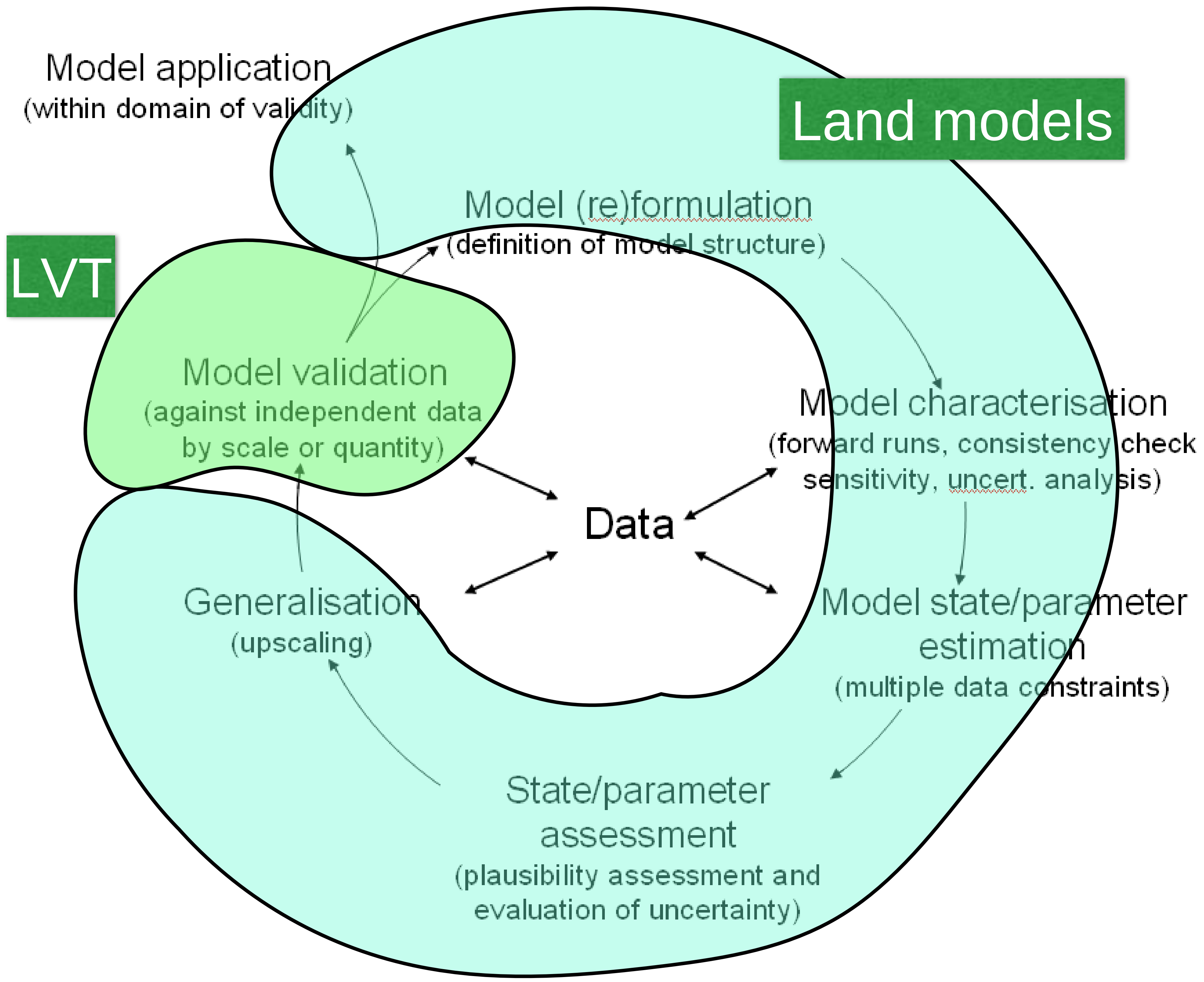
BY BART VAN DEN HURK, MARTIN BEST, PAUL DIRMEYER, ANDY PITMAN, JAN POLCHER, AND JOE SANTANELLO

The Global Land Atmosphere System Study has ushered in an era in which LSMs for numerical weather and climate prediction now incorporate complex vegetation responses, detailed hydrology, dynamic snowpack evolution, urban processes, and more.

Need a general benchmarking framework capable of capturing useful modes of variability of LSMs through a range of performance metrics is necessary for further advancing the performance and predictability of models

- Quantitative measures of fidelity of model simulations are essential for improving the usage and acceptability of model forecasts for real-world applications
- Characterization of accuracy and uncertainty in model predictions - to be used as a benchmark for future model enhancements

Model-Data-Fusion (MDF)

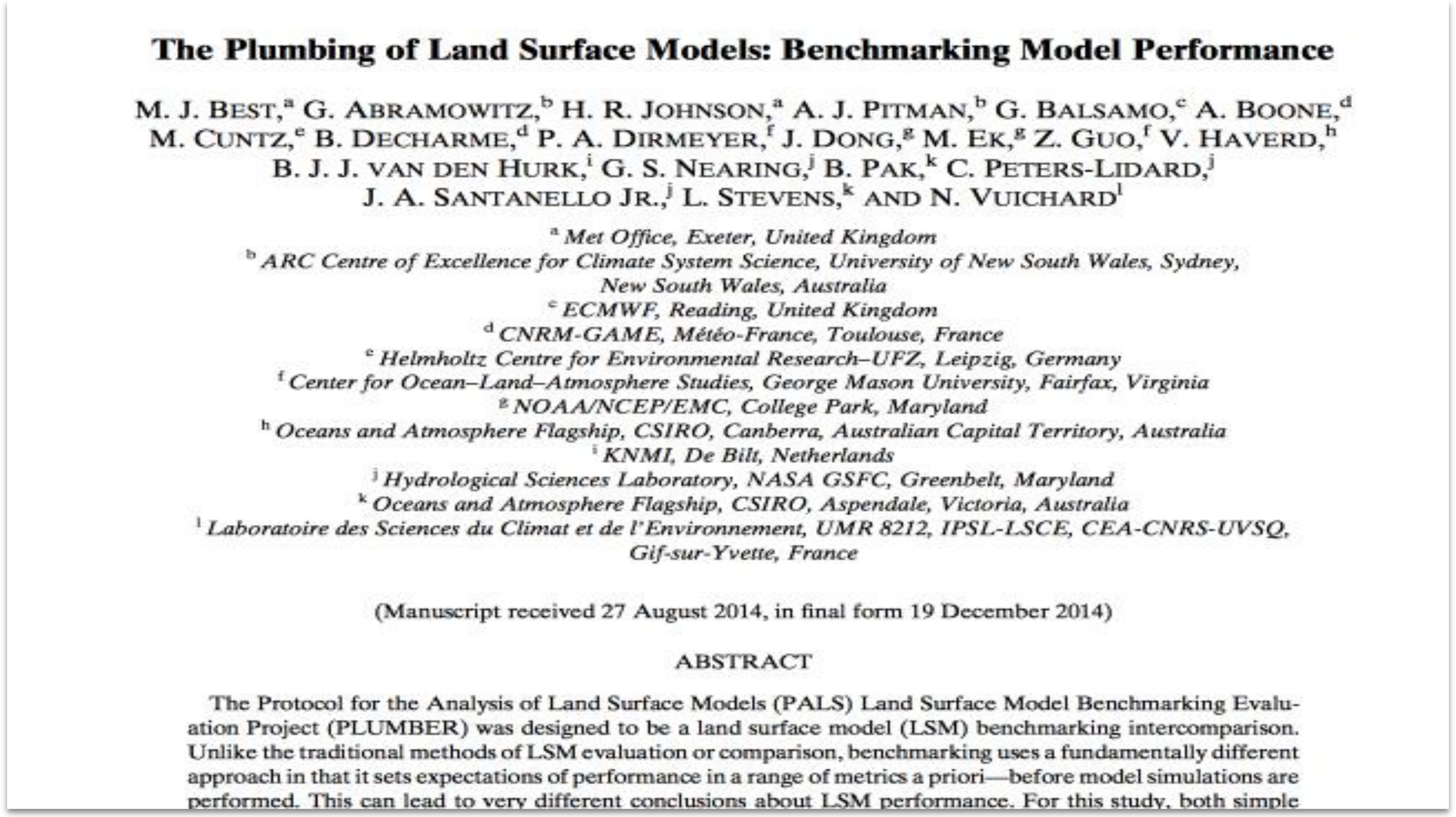


Willliams et al. (2009)

- MDF - the paradigm for combining information from models and data
- Use the information from data to help to formulation, characterization and evaluation of models in a structured manner
- MDF and Benchmarking are two of the core themes of the GEWEX GLASS community
- A comprehensive evaluation and benchmarking framework is essential for enabling the MDF concept

Definitions

LVT functions both as a **verification** and **benchmarking** environment


The Plumbing of Land Surface Models: Benchmarking Model Performance

M. J. Best,[a] G. Abramowitz,[b] H. R. Johnson,[a] A. J. Pitman,[b] G. Balsamo,[c] A. Boone,[d] M. Cuntz,[e] B. Decharme,[d] P. A. Dirmeyer,[f] J. Dong,[g] M. Ek,[g] Z. Guo,[f] V. Haverd,[h] B. J. J. van den Hurk,[i] G. S. Nearing,[j] B. Pak,[k] C. Peters-Lidard,[j] J. A. Santanello Jr.,[j] L. Stevens,[k] and N. Vuichard[l]

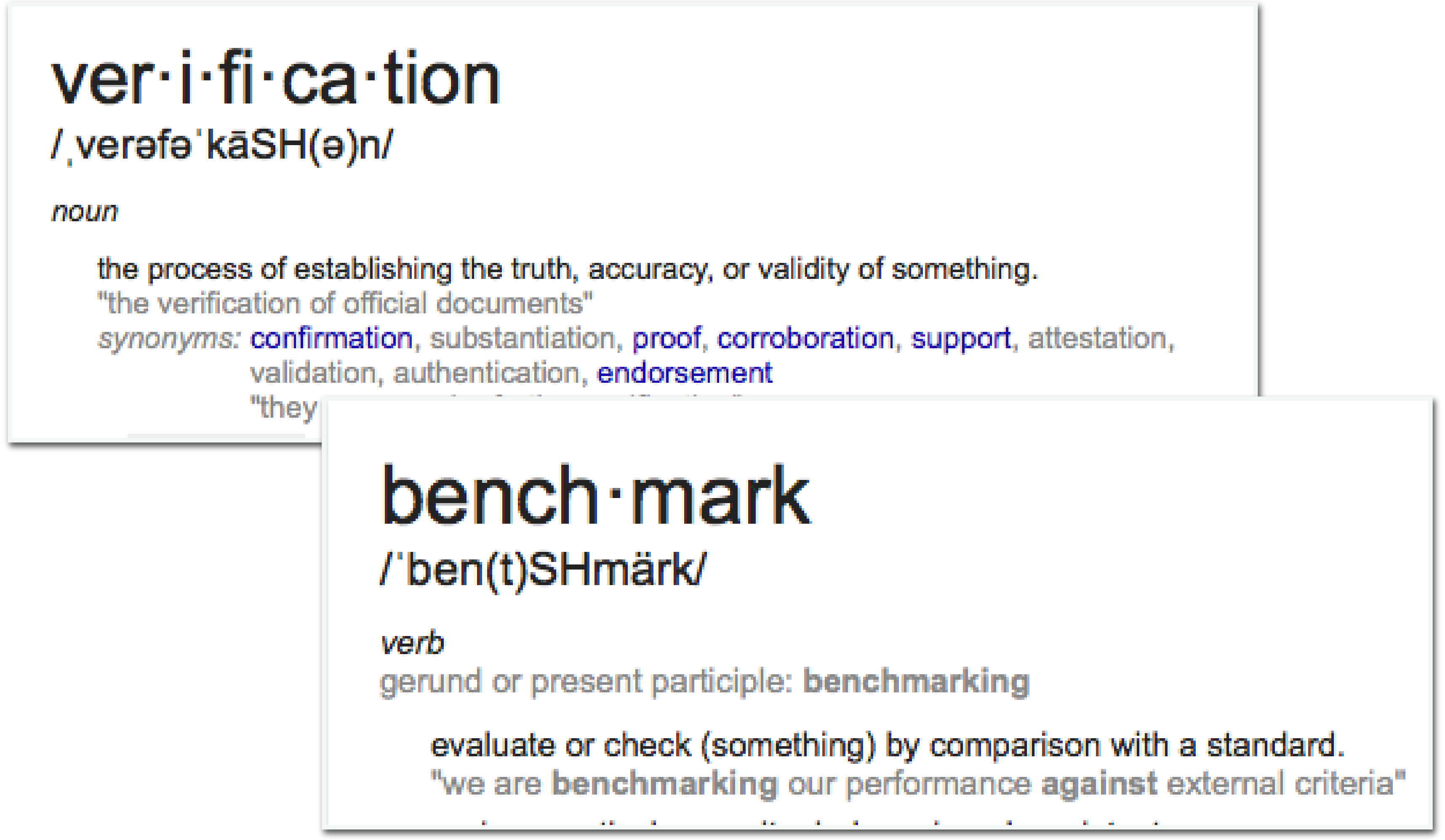
[a] Met Office, Exeter, United Kingdom
[b] ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia
[c] ECMWF, Reading, United Kingdom
[d] CNRM-GAME, Météo-France, Toulouse, France
[e] Helmholtz Centre for Environmental Research–UFZ, Leipzig, Germany
[f] Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia
[g] NOAA/NCEP/EMC, College Park, Maryland
[h] Oceans and Atmosphere Flagship, CSIRO, Canberra, Australian Capital Territory, Australia
[i] KNMI, De Bilt, Netherlands
[j] Hydrological Sciences Laboratory, NASA GSFC, Greenbelt, Maryland
[k] Oceans and Atmosphere Flagship, CSIRO, Aspendale, Victoria, Australia
[l] Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, IPSL-LSCE, CEA-CNRS-UVSQ, Gif-sur-Yvette, France

(Manuscript received 27 August 2014, in final form 19 December 2014)

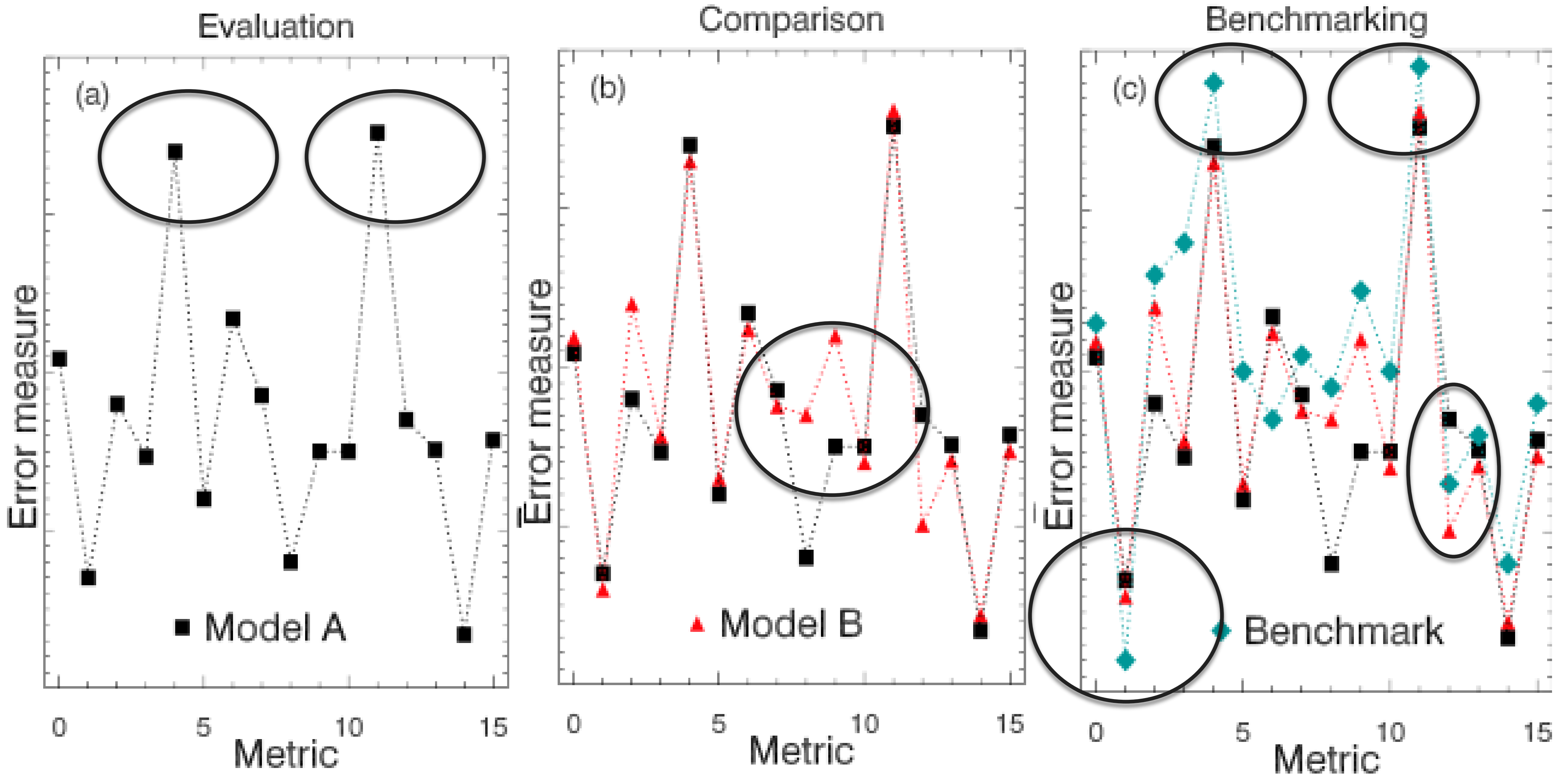
ABSTRACT

The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) was designed to be a land surface model (LSM) benchmarking intercomparison. Unlike the traditional methods of LSM evaluation or comparison, benchmarking uses a fundamentally different approach in that it sets expectations of performance in a range of metrics a priori—before model simulations are performed. This can lead to very different conclusions about LSM performance. For this study, both simple

- **Evaluation** - model outputs are compared to observations to derive an error measure
- **Comparison** - model is not just compared to observations, but also to other models
- **Benchmarking** - performance expectation defined a priori

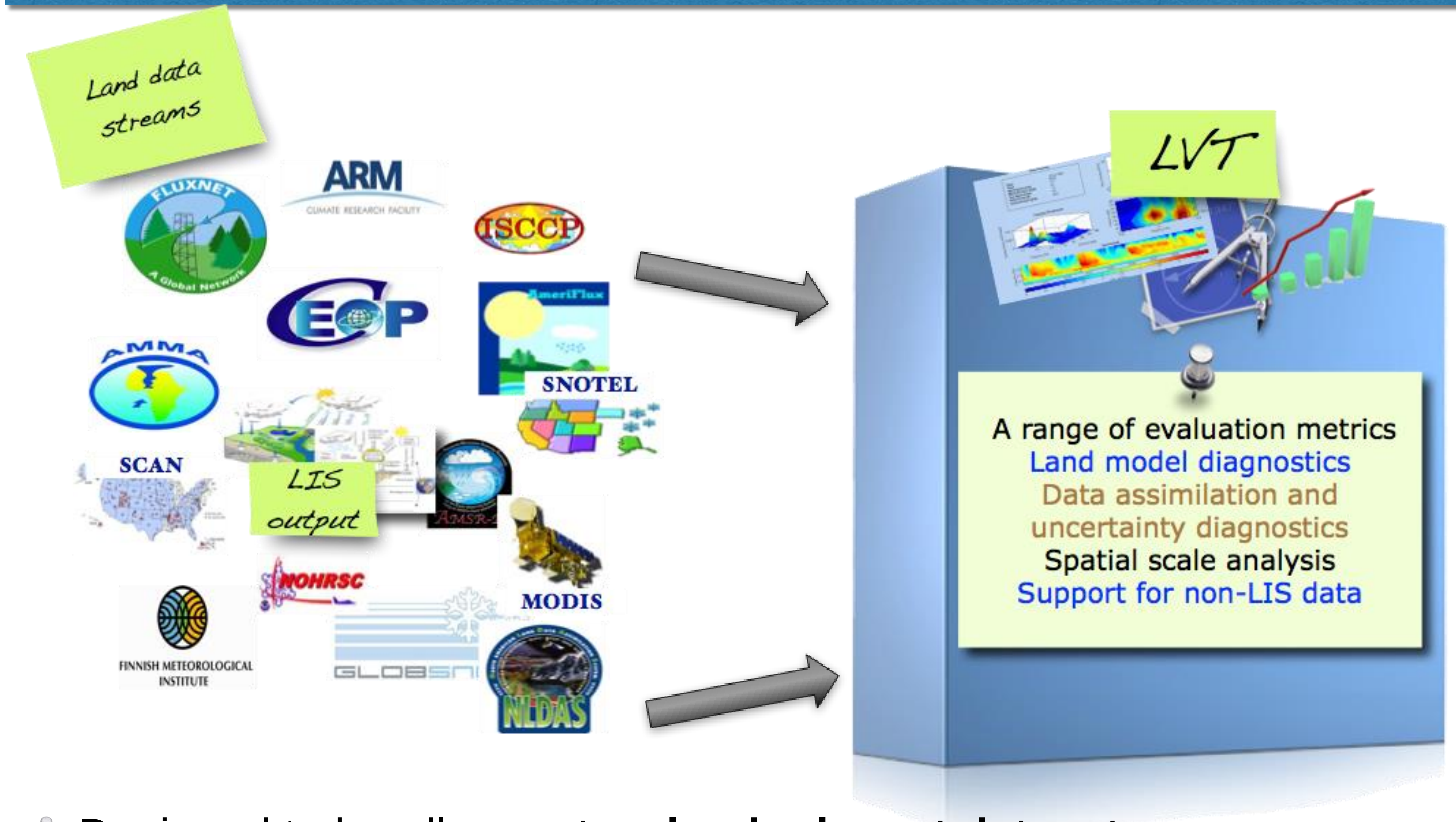

ver·i·fi·ca·tion
/ˌverəfəˈkāSH(ə)n/
noun
the process of establishing the truth, accuracy, or validity of something.
"the verification of official documents"
synonyms: confirmation, substantiation, proof, corroboration, support, attestation, validation, authentication, endorsement
"they

bench·mark
/ˈben(t)SHmärk/
verb
gerund or present participle: **benchmarking**
evaluate or check (something) by comparison with a standard.
"we are **benchmarking** our performance **against** external criteria"



source: Best et al. (2015)

LVT - overview of current capabilities



- **Open source software**
- **Includes support for a range of in-situ, remote sensing and model/renalysis products**
- **Supports a range of metrics (diagnostics, deterministic, information-theory, decision-theory, scale-decomposition based metrics)**
- **Includes the capability to generate end-user oriented hydrological products (drought/flood percentiles, indicators)**

- Designed to handle **any two land relevant datasets**
- Supports benchmarking using methods from linear regression to more complex methods.
- The supported datasets in LVT can be used to develop benchmarks

Kumar et al. (2012), “Land surface Verification Toolkit (LVT) - a generalized framework for land surface model evaluation, Geosci. Model Dev., 5, 869-886.

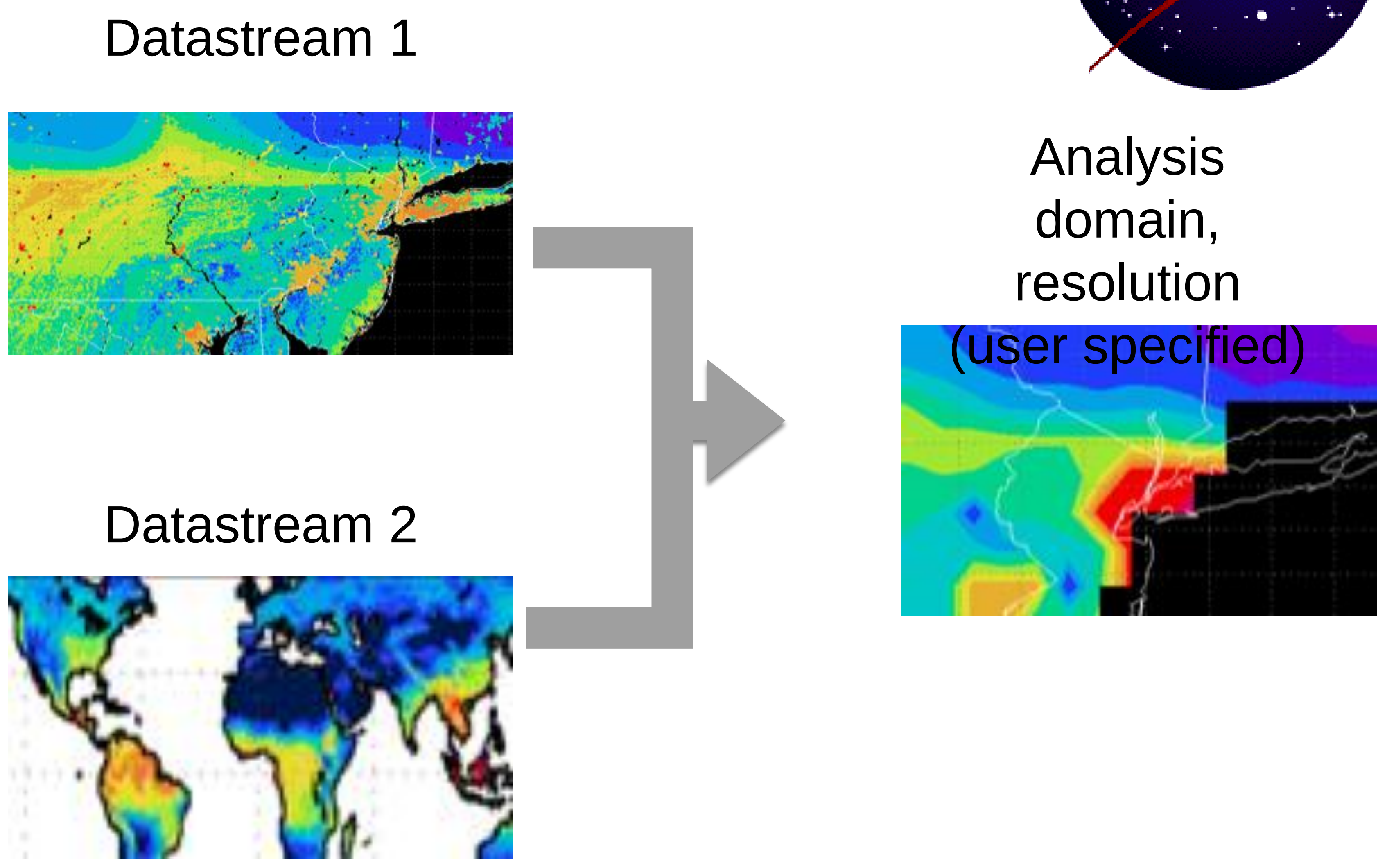
General capabilities

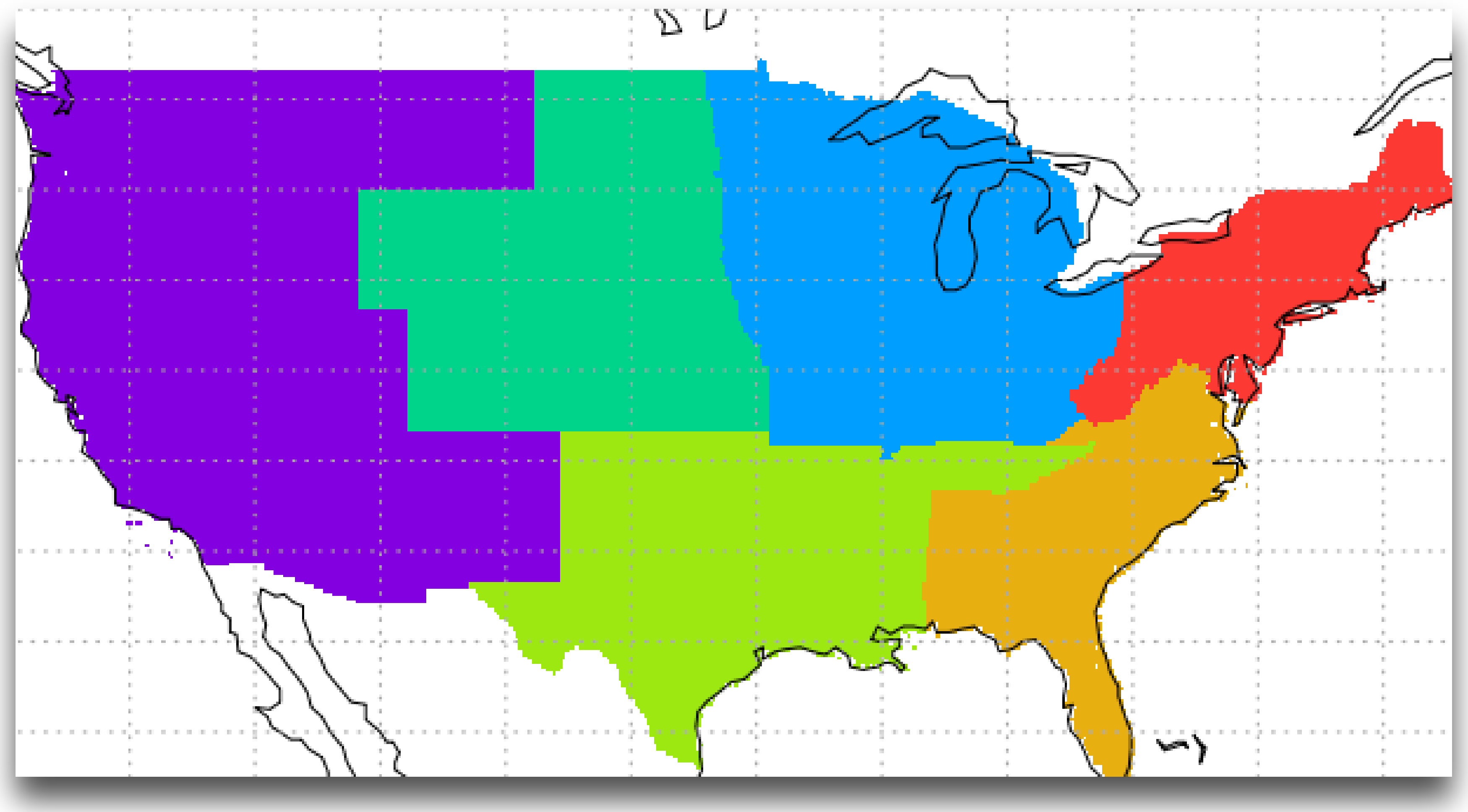
- Works with the datasets natively and reconciles the differences in spatial and temporal resolutions between the two datastreams being compared, by bringing them to a common (user specified space and time domain)

- Support for data masking, spatial, temporal stratification, time-lagged computations

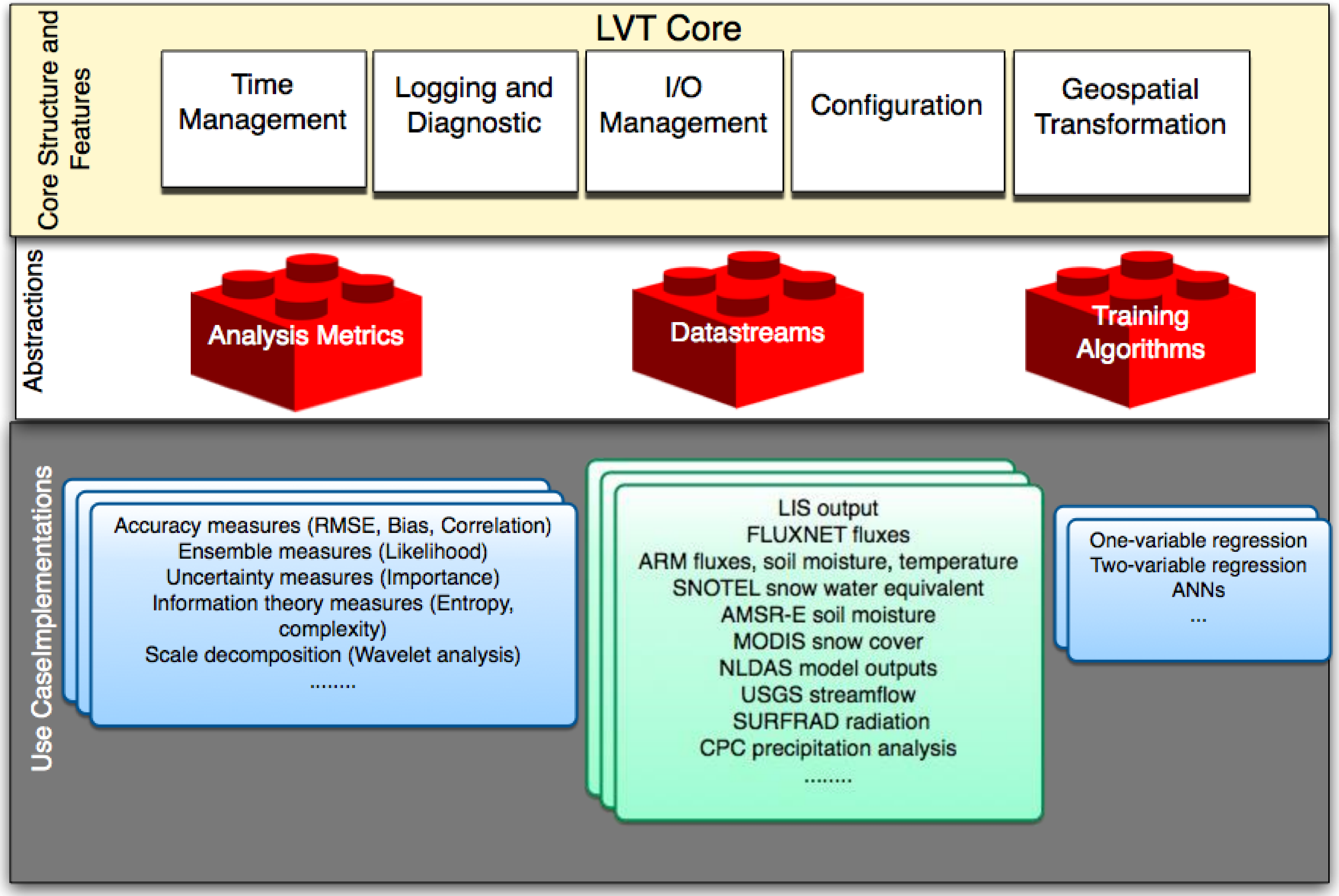
- Metric computations supported on a grid cell by grid cell basis or at basin scales.

- Computes confidence intervals on the metric calculations; Calculates derived variables (e.g. bowen ratio)



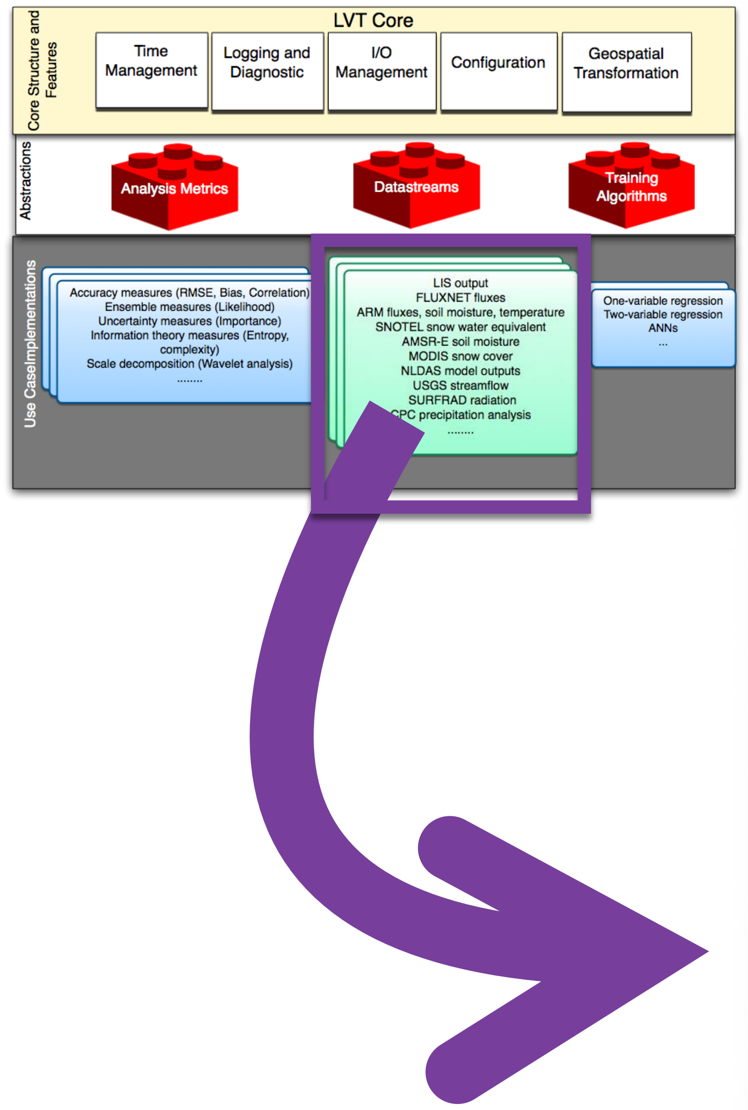


Software architecture



- 3-layer architecture
- Specified as an object oriented framework with plugins defined for
  - Analysis metrics
  - Datastreams
  - Training algorithms
- Analysis instances are enabled by a config file (no external scripting required)

Supported data streams



In-situ

- Ameriflux fluxes
- ARM fluxes, soil moisture, soil temperature
- ARS soil moisture
- CEOP fluxes, soil moisture, soil temperature
- CPC precipitation
- FLUXNET fluxes
- FMI SWE
- GHCN snow depth
- GLERL lake fluxes, temperature
- ISMN soil moisture
- NASMD soil moisture
- PBOH2O soil moisture, snow depth
- SCAN soil moisture
- SMOSREX soil moisture
- SNODEP snow depth metobs
- SNOTEL SWE
- SURFRAD radiation
- USGS streamflow
- USGS groundwater
- ..

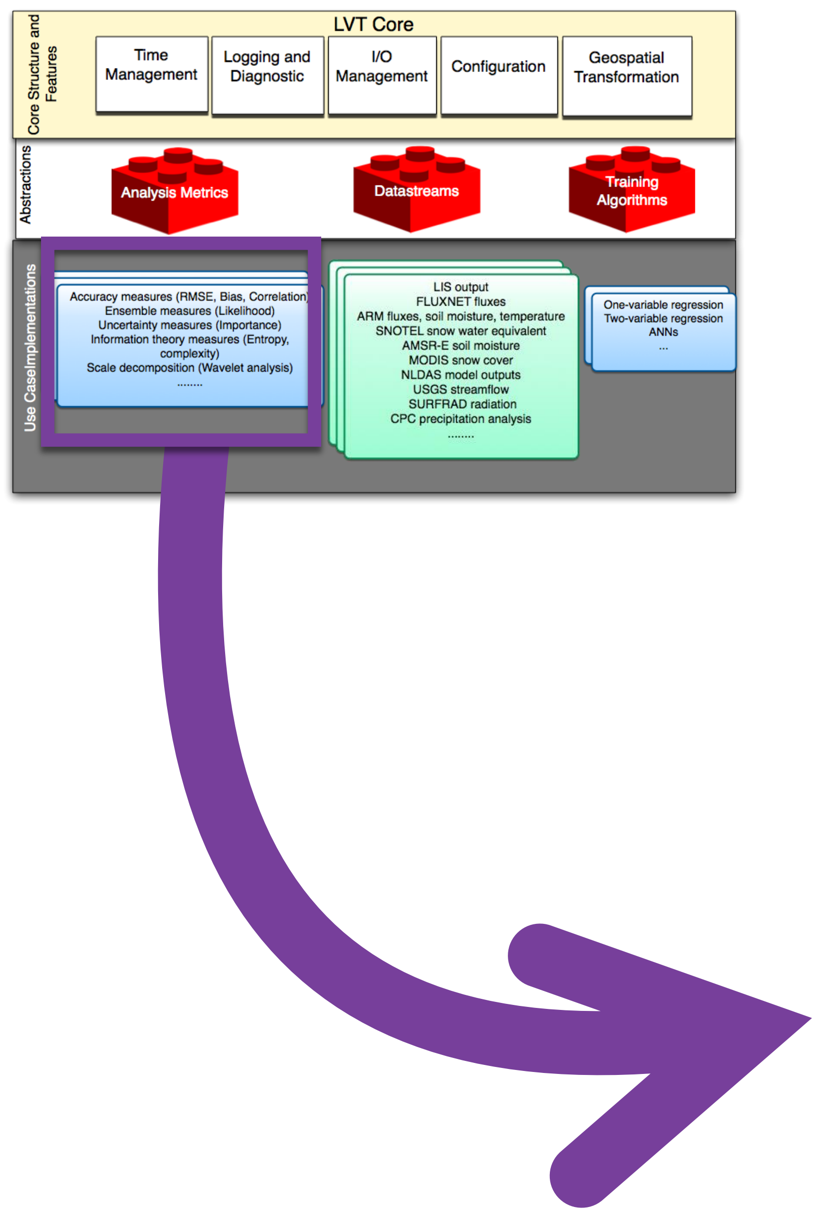
Satellite/Remote Sensing

- ALEXI
- AMSR-E SWE/snowdepth
- LPRM AMSR-E soil moisture
- ESA CCI soil moisture
- GIMMS NDVI
- MODIS NDVI
- GlobSnow SWE
- GRACE TWS
- ISCCP LST
- MOD10A1 snow cover
- MOD16A2 ET
- MODIS LST
- SMOPS soil moisture
- SMOS L1 Tb
- SMOS L2 soil moisture
- UW ET
- SMAP
- …

Model/Reanalysis

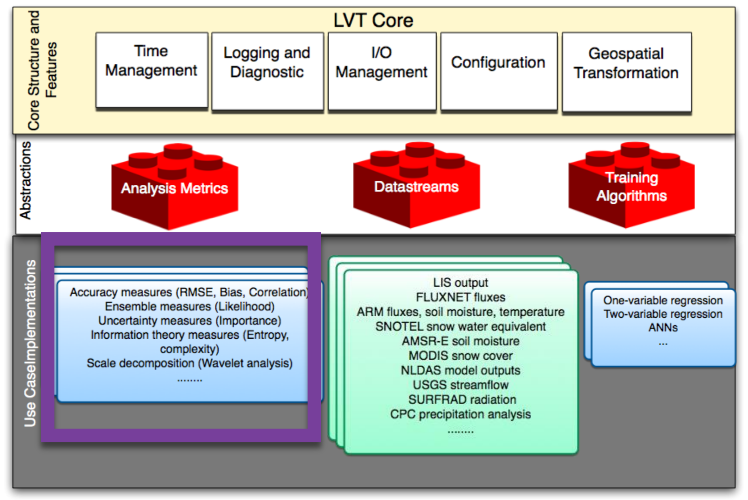
- AGRMET
- GLDAS2
- NLDAS2
- LIS outputs
- MERRA2
- SNODAS
- CMC
- GL6 JULES
- ERA interim Land
- MERRA Land
- MERRA2
- ….

Supported analysis metrics



Metric class	Examples
Diagnostics	Mean, Standard deviation, Anomaly, Tendency, Min, Max, Sum, Maxtime, Mintime
Accuracy	ACC, Bias, CSI, ETS, FAR, FBIAS, MAE, NSE, PODY, PODN, POFD, Correlation, Anomaly Correlation, Tendency Correlation, unbiased RMSE
Indicators	SPI, SRI, SSWI, SSGI, percentiles, probabilistic percentiles
Ensemble/Uncertainty	Mean, Likelihood, Spread, Cross correlation, ME
Information theory	Metric entropy, Information gain, Effective complexity, Fluctuation complexity
Scale decomposition	Discrete wavelet transforms
Spatial similarity	Hausdorff norm

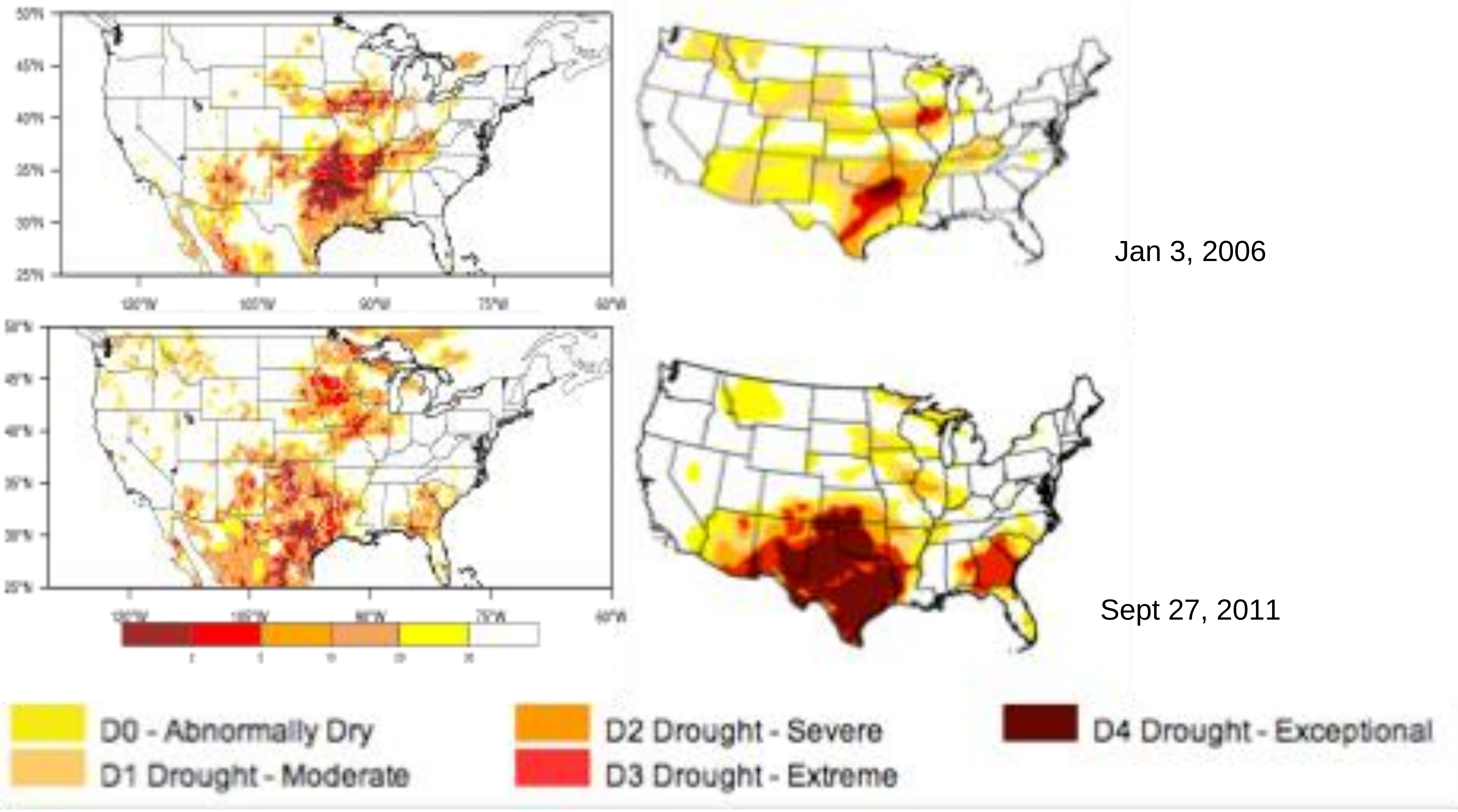
Examples of indicators (metrics of extremes)



- A suite of common, normalized indicators has been developed
  - SPI, SRI, SSWI, SSGI, percentiles
- These indicators are computed as deviations from long term (fitted/computed) distributions

Root zone soil moisture based drought percentiles generated by LVT from a LSM run

U.S. Drought Monitor estimate

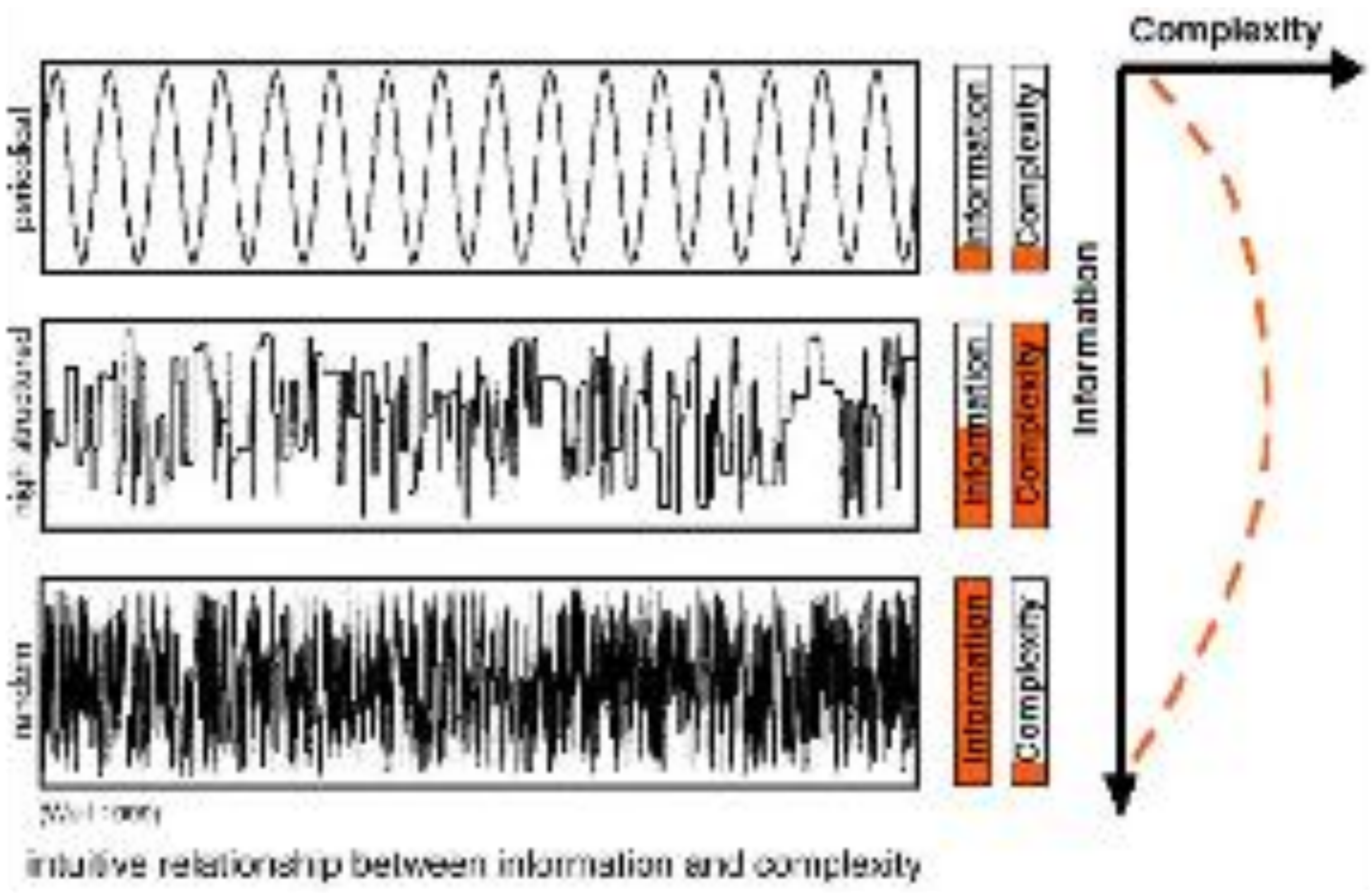


Information theory metrics

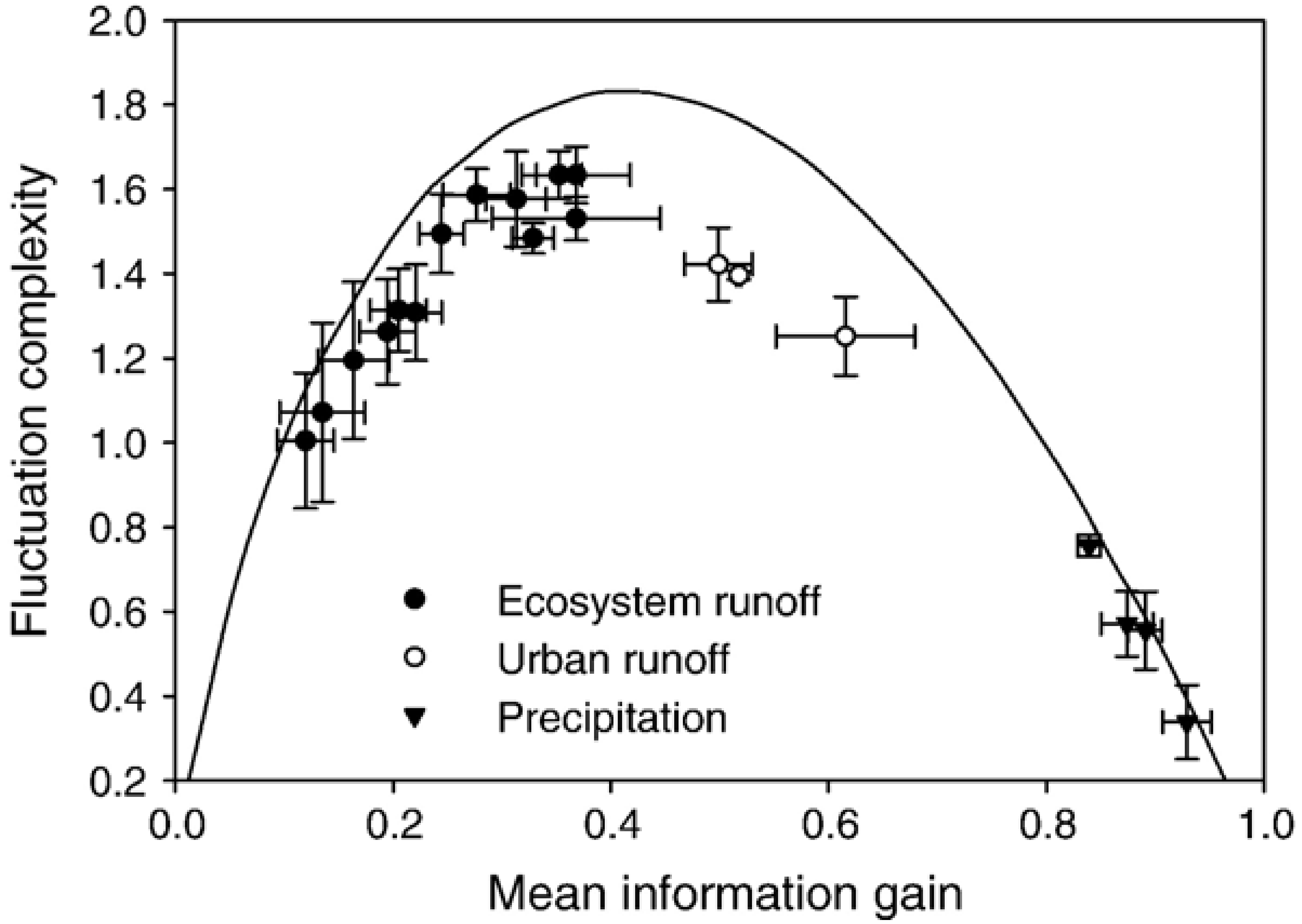
Developed by Claude Shannon to find fundamental limits on signal processing

Entropy - quantifies the uncertainty involved in predicting the value of a random variable

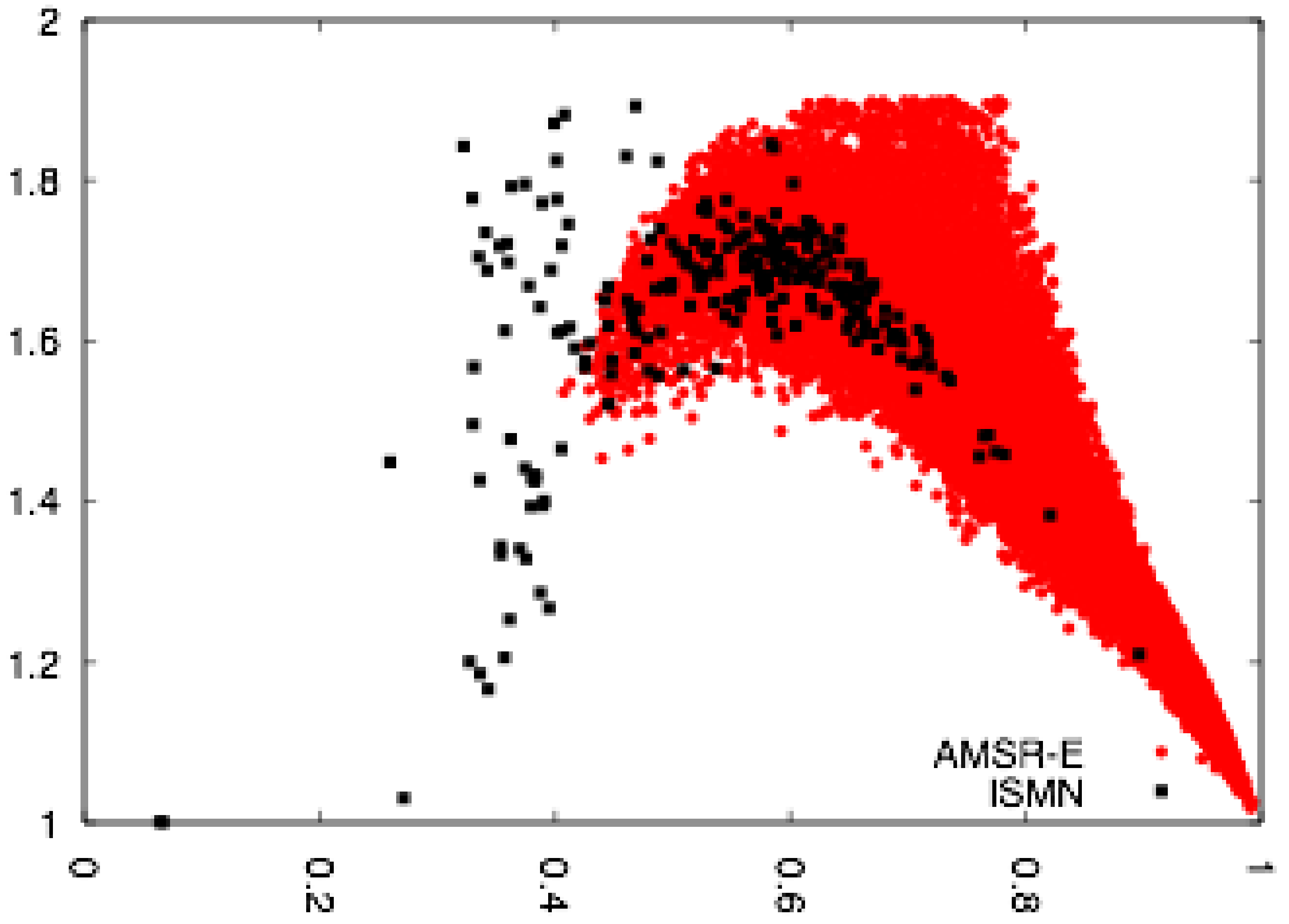
Complexity - small for periodic sequences and completely random data

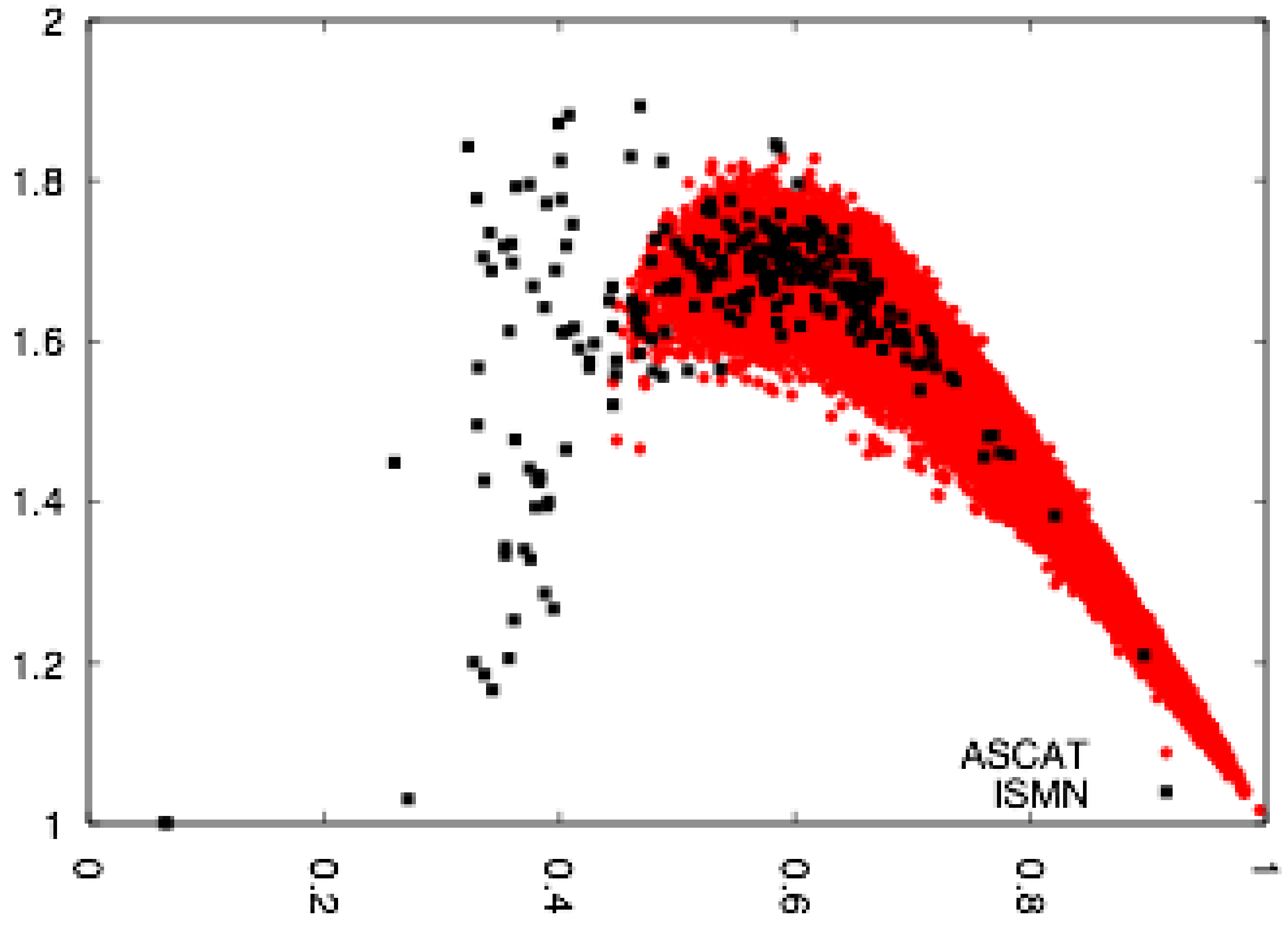


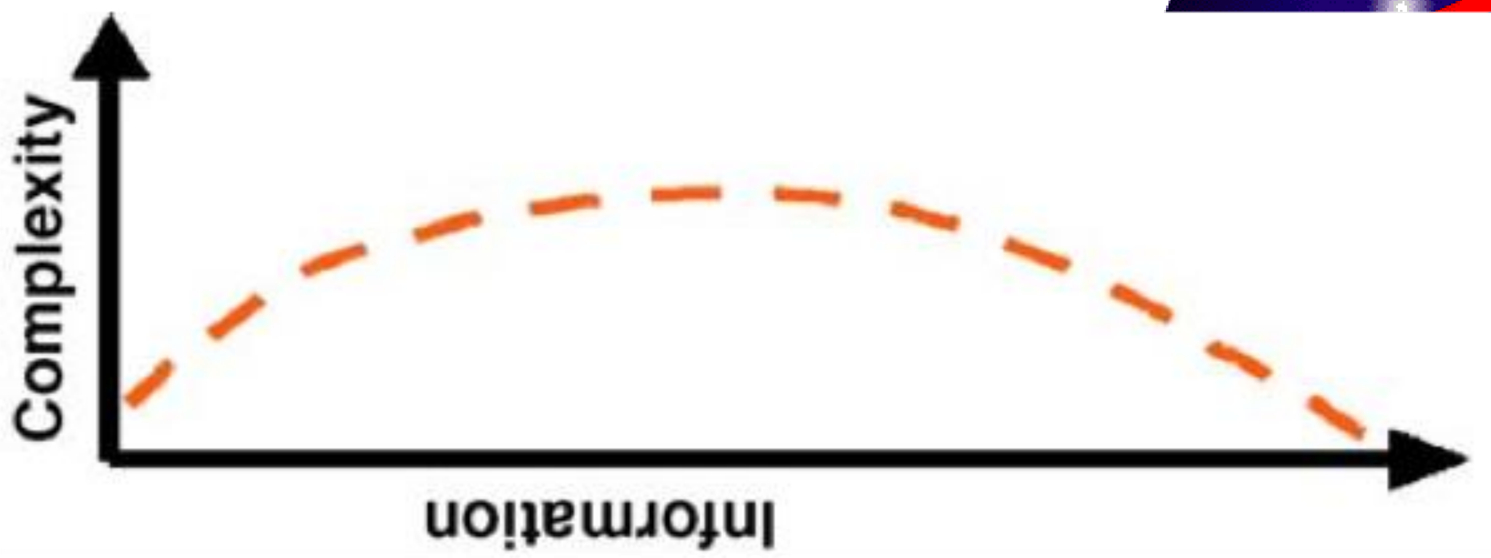
from Pachepsky et al. (2006)

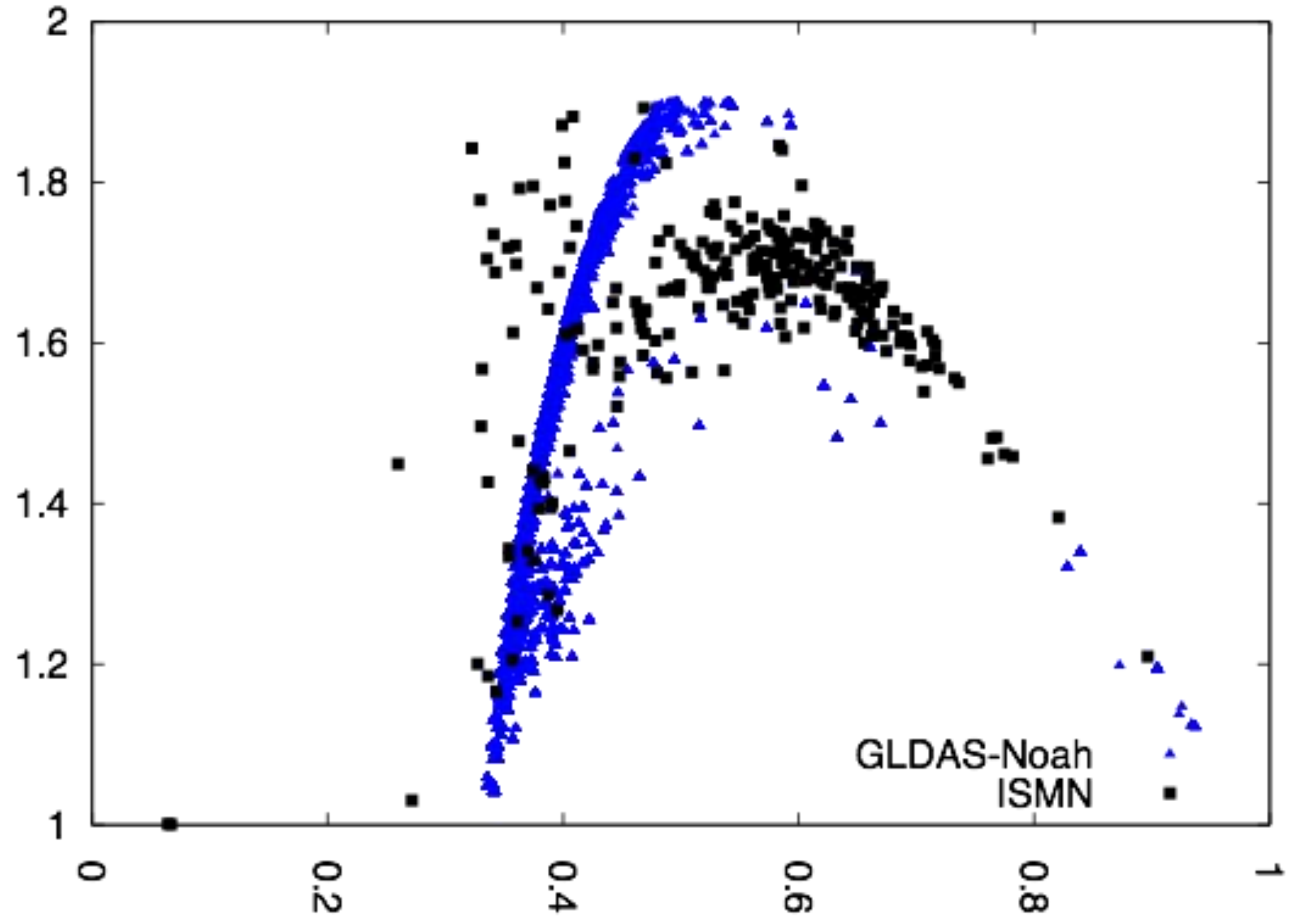


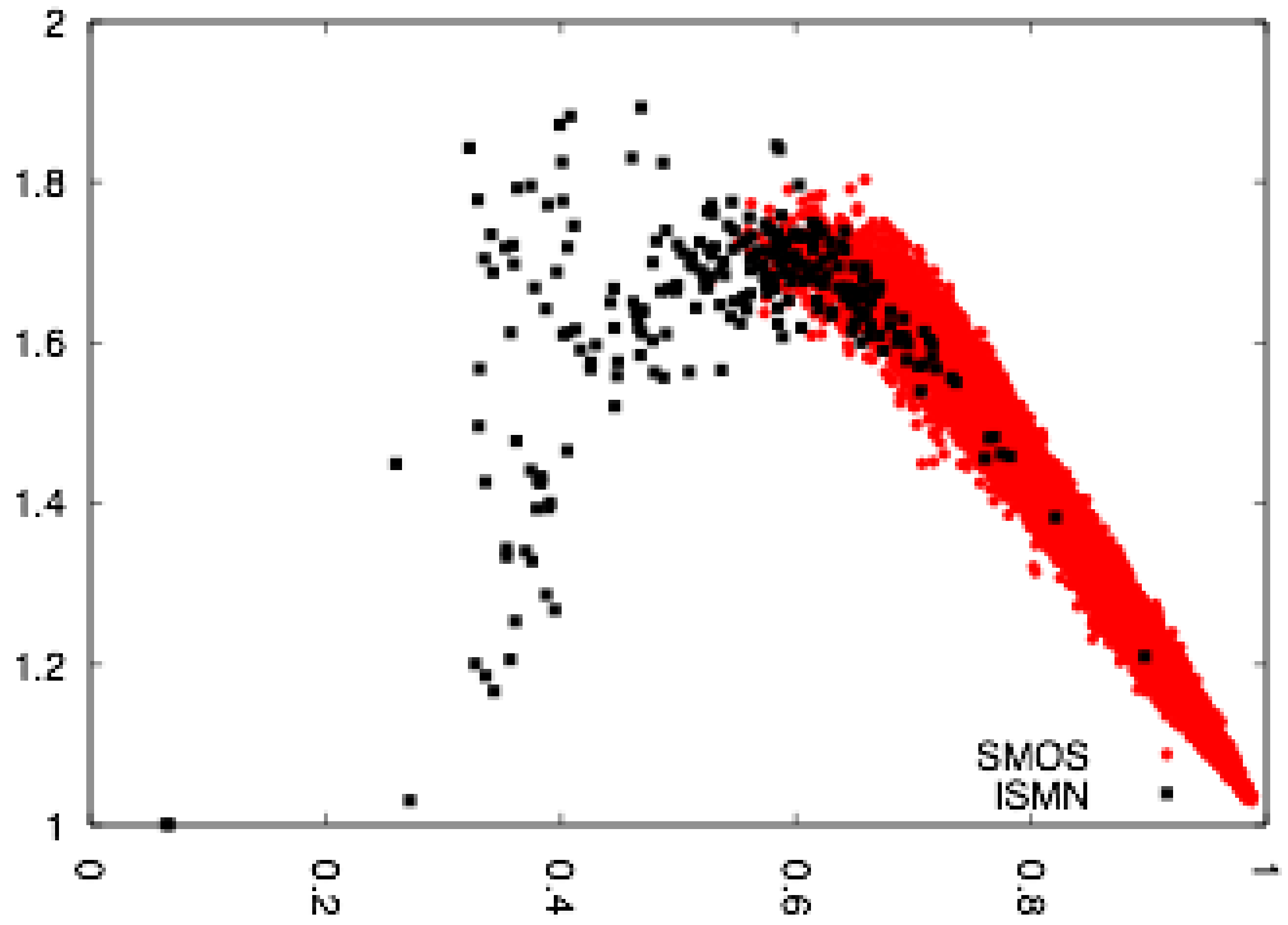
Information theory metrics application

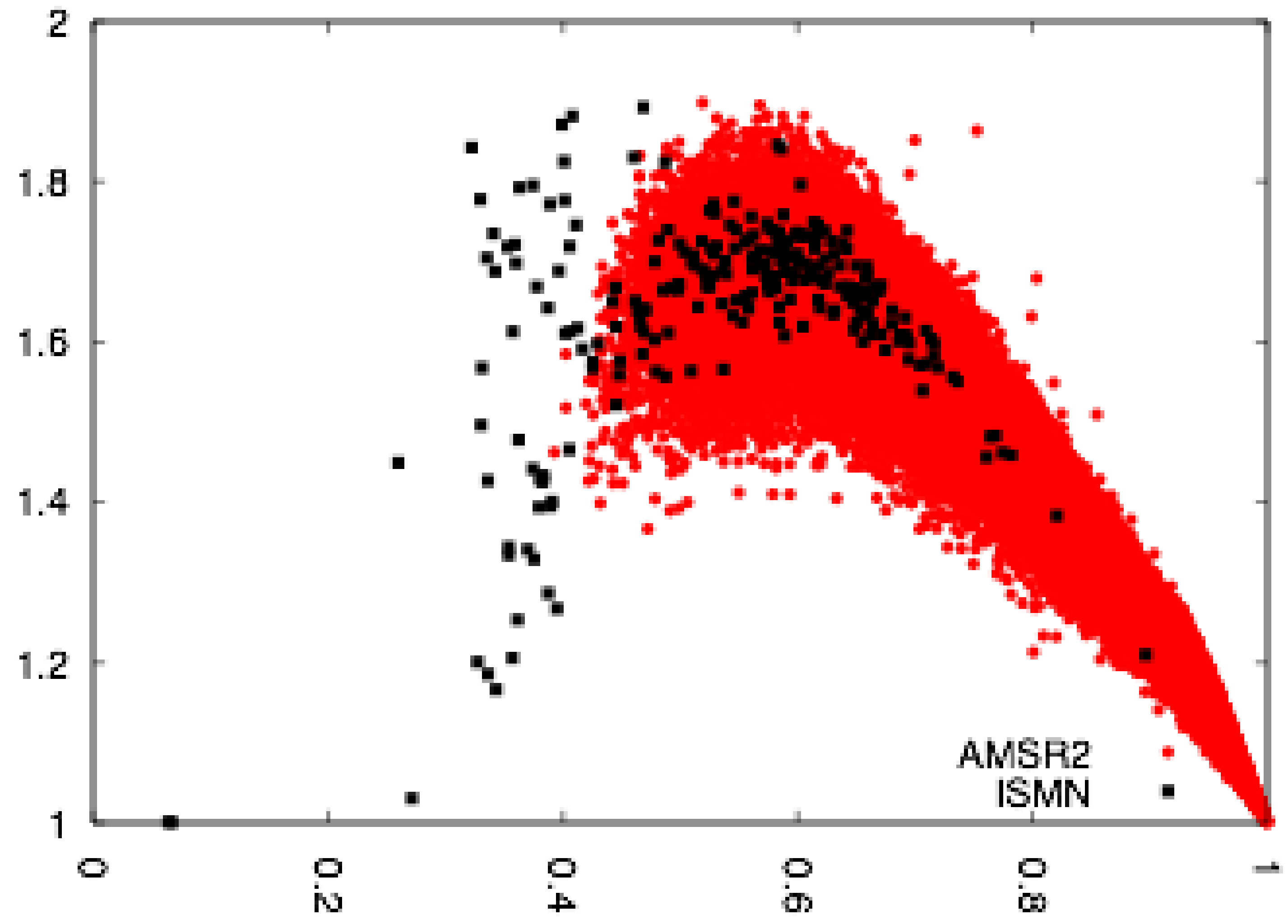


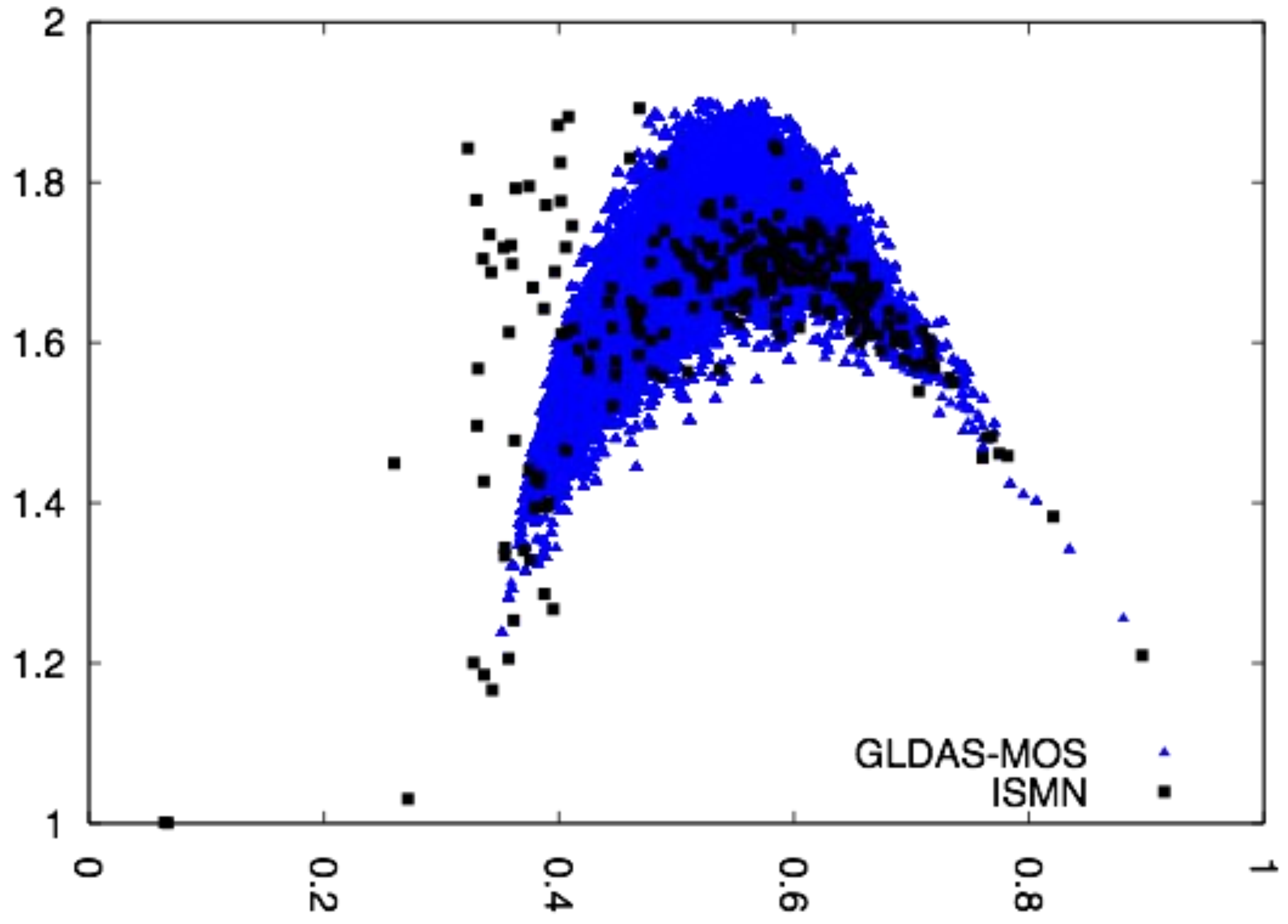




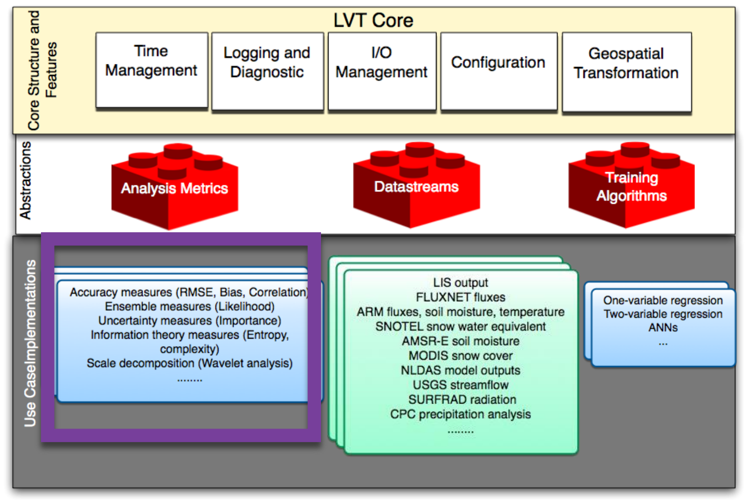


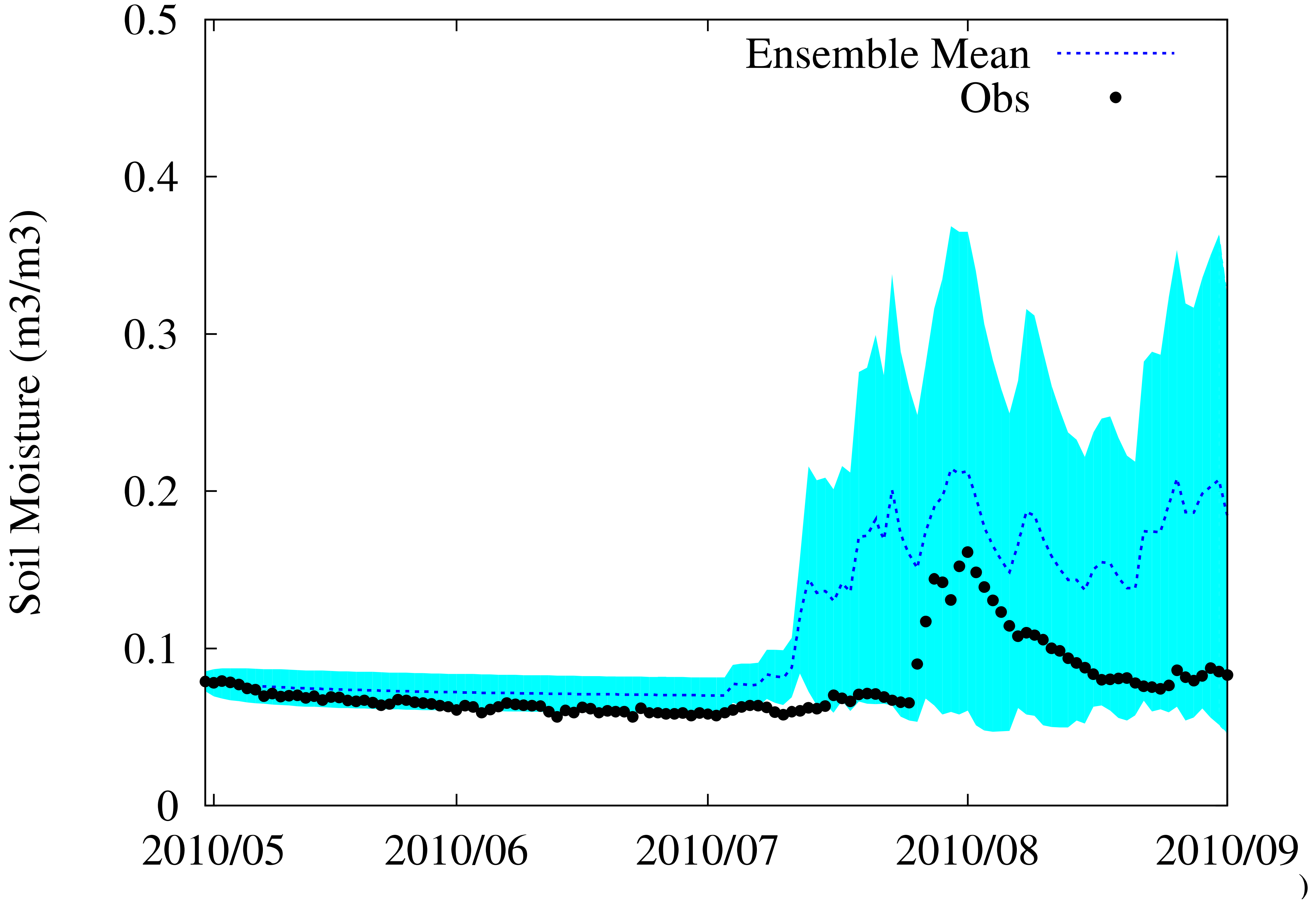


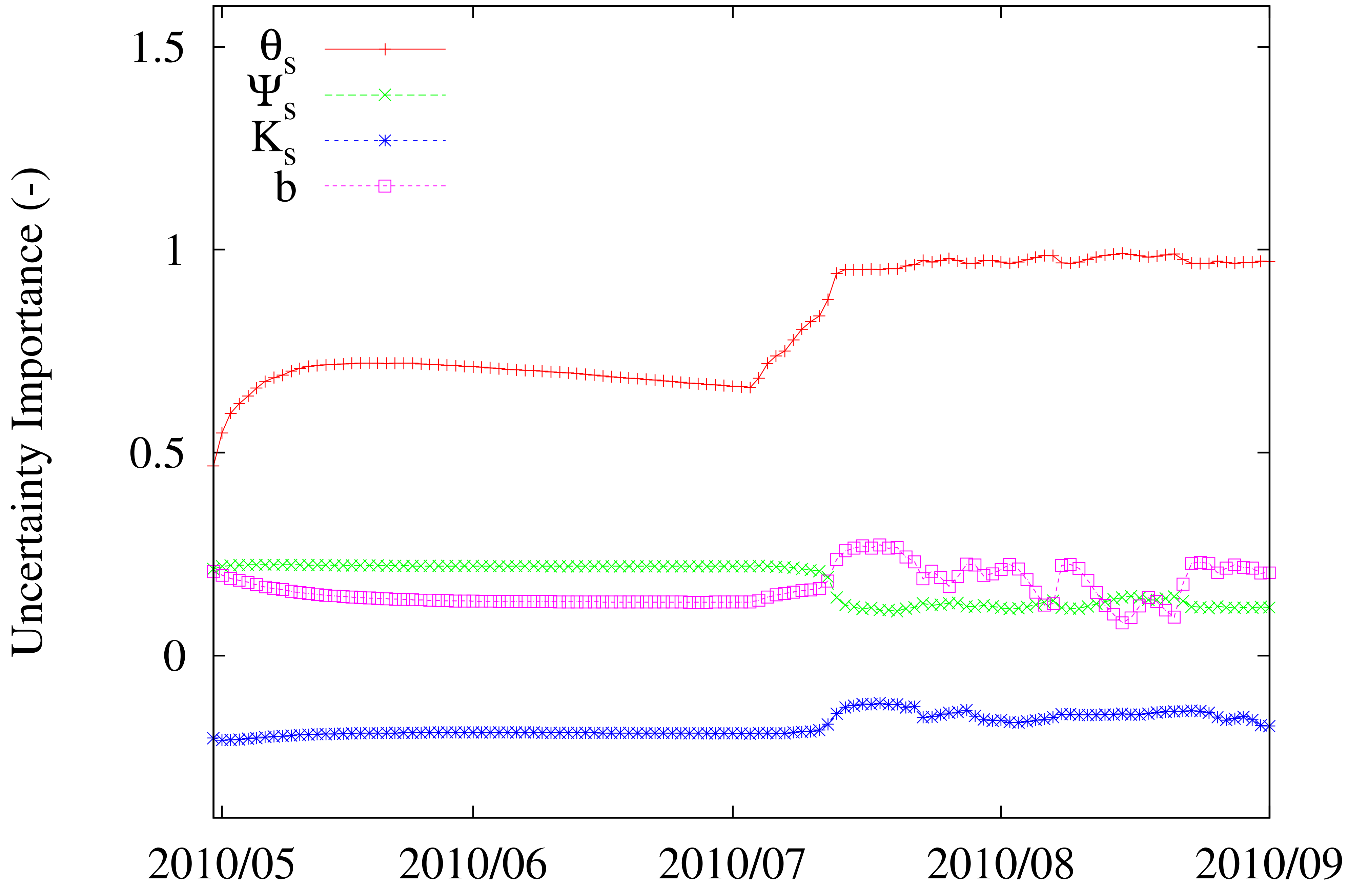




Uncertainty quantification and analysis



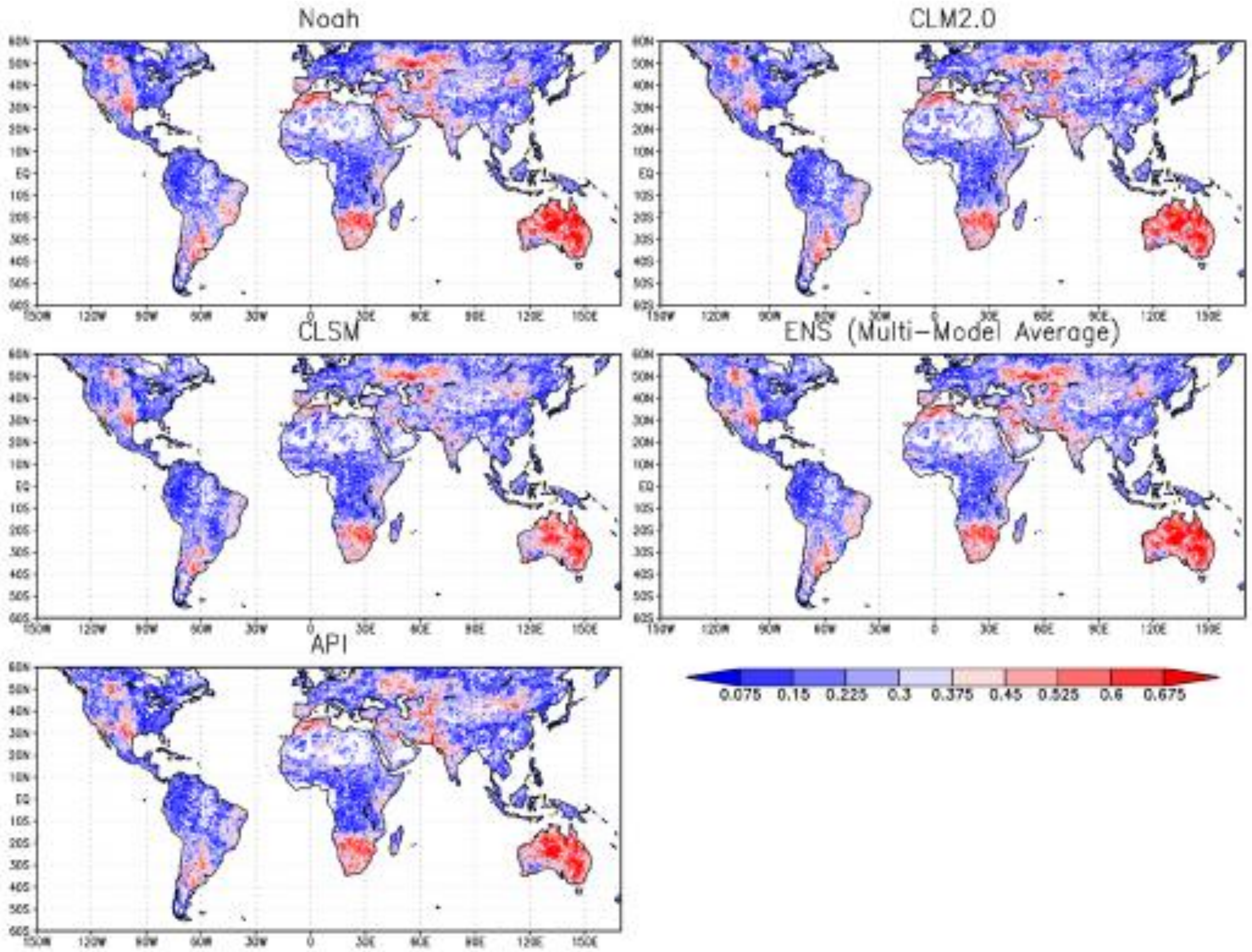




- Uncertainty importance: An assessment of the relative contribution of each parameter to the ensemble spread (cross correlation between the simulated variable and the parameter, across the ensemble)
- Can be used to guide parameter optimization/uncertainty estimation studies

Remote sensing data analysis

1 month lagged rank cross correlation of root zone soil moisture and VI



- Root zone soil moisture is a leading indicator of subsequent vegetation anomalies. Under water stressed conditions, negative soil moisture anomalies should temporally precede a detectable impact on vegetation health.
- Lagged rank cross-correlation between model-derived root-zone soil moisture estimates and remotely sensed vegetation indices (MOD13C2) from LSMs of varying complexity are compared.
- Strong coupling in semi-arid areas of the world, weak coupling in humid regions
- Not a lot of added skill in the modern land surface models compared to the API model.

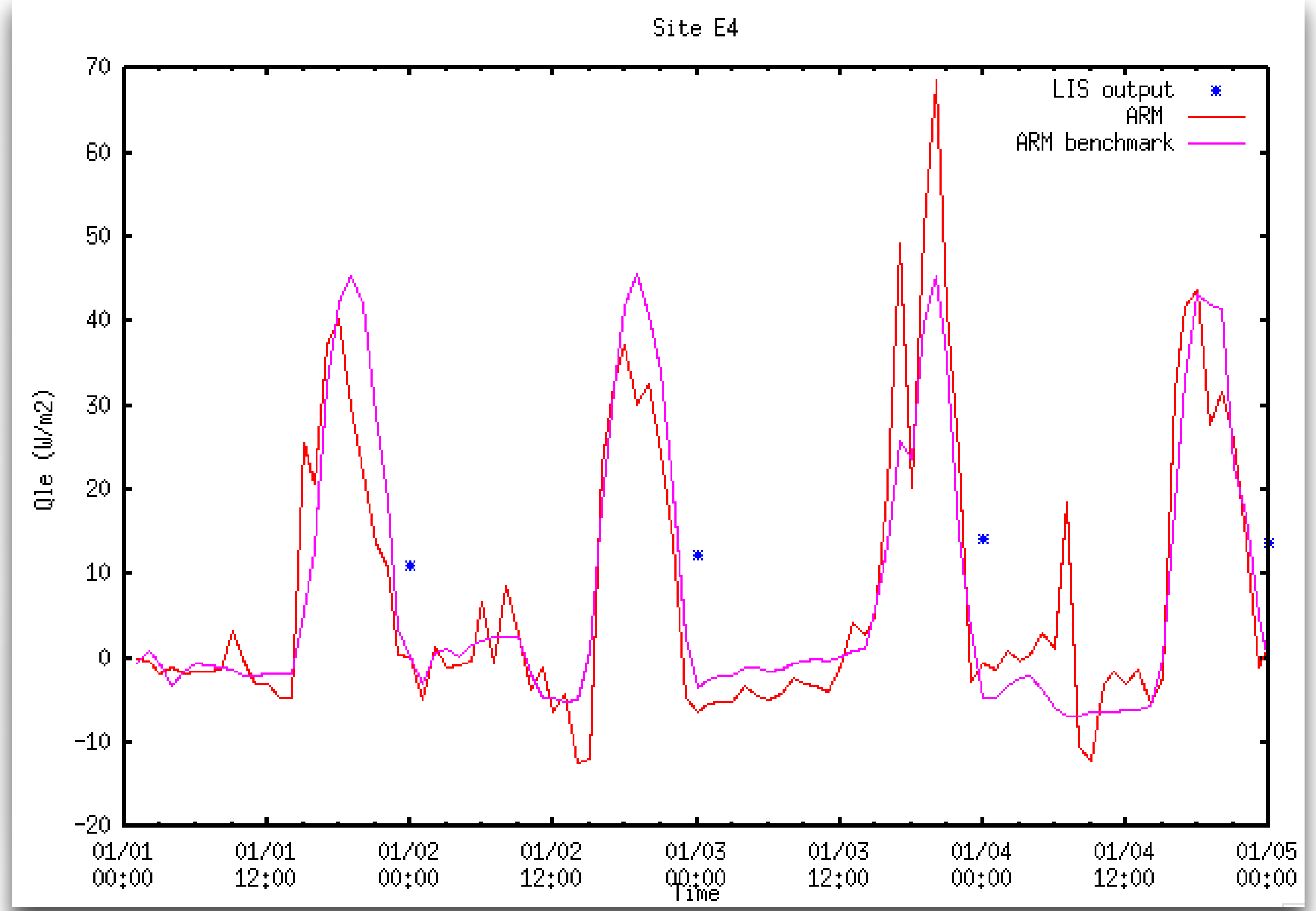
Crow et al. (2012), “On the utility of land surface models for agricultural drought monitoring”, HESS, 16, 3451-3460.

Benchmarking

- LVT provides two capabilities related to benchmarking:
  - Develop a benchmark dataset by training any two of the supported datasets
  - Compare the model runs to the benchmark dataset
- Training algorithms available (One/Two-variable regression; ANN - in development)

- Example at ARM-SGP for benchmarking latent heat flux (Qle) estimation

1. Compare the model simulation of Qle vs. ARM data
2. Develop a benchmark using LVT by training ARM Qle measurements to ARM net radiation and air temperature measurements (using out of sample, two-variable regression)
3. Compare the model simulation of Qle vs benchmark



Summary

- LVT - an open source, formal model evaluation and benchmarking environment
- Facilitates the integration of in-situ, remote sensing and model/renalysis data products.
- Supports deterministic/probabilistic evaluation, a wide variety of traditional and non-traditional metrics and benchmarking strategies
- Supports hydrologic variables primarily; easy to extend the support for ecosystem/biogeochemical variables.
- Ongoing work include support for newer datasets, spectral/cross-spectral analysis, data assimilation diagnostics
- AMS land surface benchmarking session – ideas/thoughts?