A metrics framework to evaluate new approaches for drought monitoring and prediction

Andy Wood
NCAR Research Applications Laboratory

collaborators:
DTF co-leads –
A. Marrioti, S. Schubert, K. Mo, C. Peters-Lidard
Jin Huang
DTF Metrics Team

US CLIVAR Summit
Virtual – July 9, 2014

NCAR UCAR

Topics

- MAPP Drought Task Force Overview
  - Objectives
  - Organization
  - Accomplishments
- Framework for Drought Science Evaluation
- Application Example
  - Streamflow Prediction
- Future Directions

MAPP Drought Task Force Overview

MAPP DTF Mission

- Bring together and facilitate MAPP-funded research efforts aimed at achieving advances in capabilities to monitor and predict drought over North America
  - *assess current efforts to advance monitoring and prediction*

- Contribute to efforts to advance official national drought products including:
  - development of a DEWS by NIDIS
  - drought monitoring/prediction activities at NCEP

- Coordinate with and leverage other relevant national and international efforts (e.g. NMME, WCRP Drought Interest Group).

MAPP Drought Task Force Overview

Key Activities

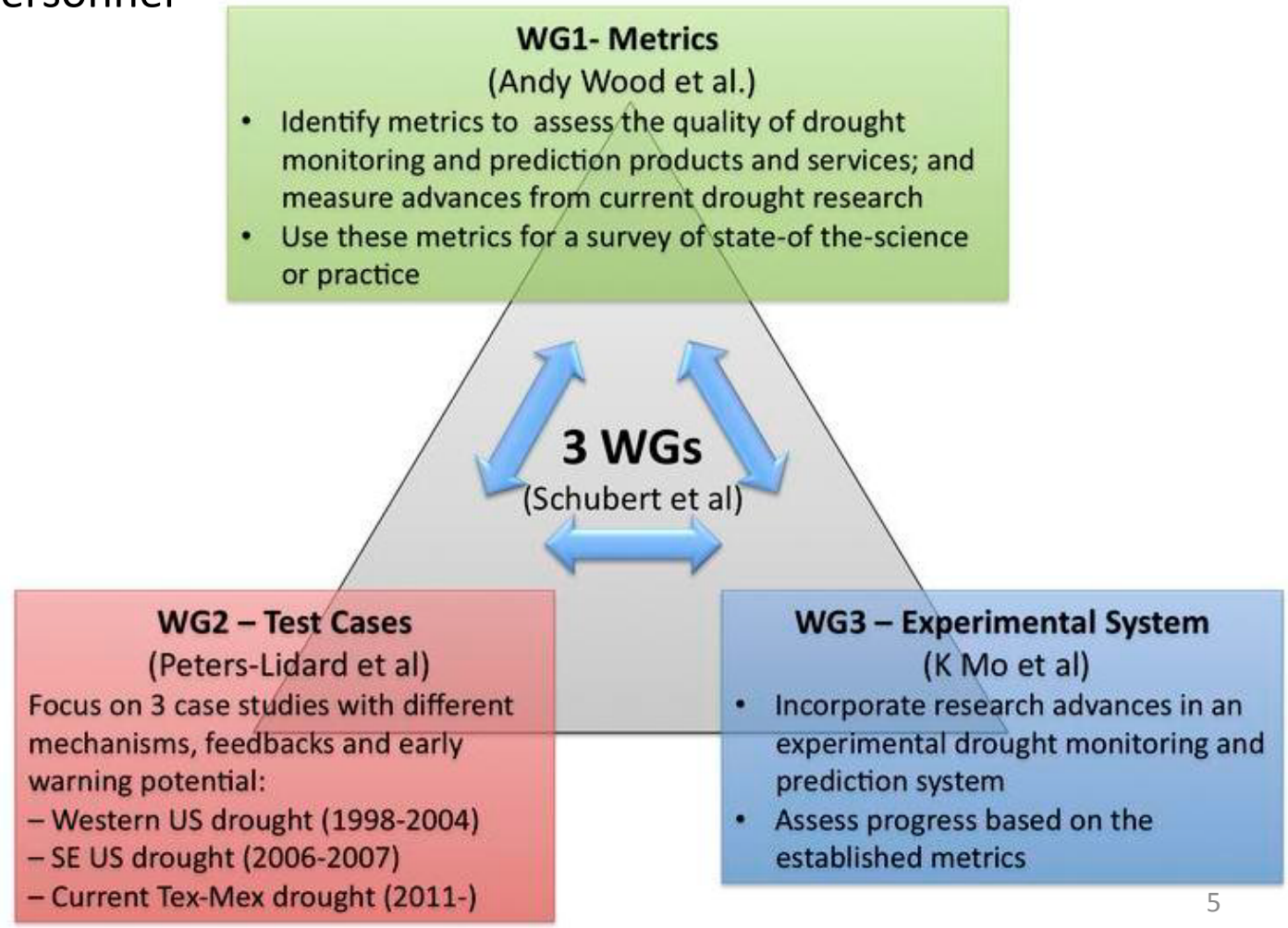
- Communication/coordination:
  - Monthly telecons
    - *Narrative, Research-to-Capability, and Science sessions*
  - DTF Wiki page (internal) and NIDIS-DTF Web Page
    http://www.drought.gov/drought/content/resources-research/noaa-drought-task-force
  - Linked with the National Multi-Model Ensemble (NMME) effort

- Organizing a special collection in JHM on "Advancing Drought Monitoring and Prediction"
  - About 16 contributed papers plus 1 overview/synthesis papers
  - http://journals.ametsoc.org/page/droughtMonitoring

- Developed DTF group working "infrastructure" – organizes research into a de facto *testbed*

MAPP Drought Task Force Overview

- Led by Annarita Mariotti (NOAA) and Siegfried Schubert (NASA)
- Supported by efforts of 35+ researchers and partner organization personnel



Framework for Drought Science Evaluation

Working Group 1 – Metrics*

- The Drought Task force is centrally focused on evaluating geophysical science supporting drought management
- Most variables of interest are physical, but index/category variables are included in focus

Drought Variables of Interest	Role(s)
Streamflow	Hydrological Drought
Precipitation	Meteorological Drought
Soil Moisture	Agricultural Drought
Snow Water Equivalent	Indirect / Multiple
Temperature	Indirect / Multiple
ET/PE	Indirect / Multiple
DM/SDO Category	Decision Support

* Note, not 'drought metrics' but 'drought *research* metrics'

Drought Capability Assessment Protocol

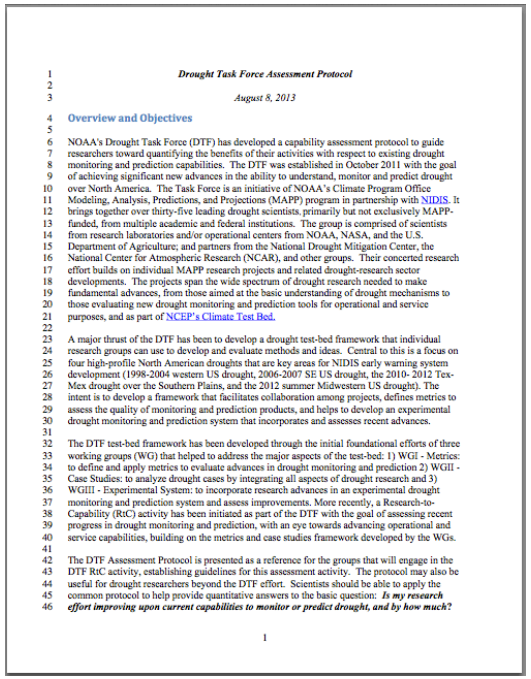
Principles

The protocol should include research performance measures that are:

- Specific to drought and define thresholds or criteria that separate drought conditions from other system states and phenomena

- Describe key geophysical features of drought that are of interest to decision makers in applications sectors and motivated by societal impacts. Examples include the onset, severity, duration, and change in intensity of a drought variable.

- Centered on the drought event case studies selected by the DTF and include the application of statistically robust metrics.



Drought Task Force Assessment Protocol

August 8, 2013

Overview and Objectives

NOAA's Drought Task Force (DTF) has developed a capability assessment protocol to guide researchers toward quantifying the benefits of their activities with respect to existing drought monitoring and prediction capabilities. The DTF was established in October 2011 with the goal of achieving significant new advances in the ability to understand, monitor and predict drought over North America. The Task Force is an initiative of NOAA's Climate Program Office Modeling, Analysis, Predictions, and Projections (MAPP) program in partnership with NIDIS. It brings together over thirty-five leading drought scientists, primarily but not exclusively MAPP-funded, from multiple academic and federal institutions. The group is comprised of scientists from research laboratories and/or operational centers from NOAA, NASA, and the U.S. Department of Agriculture; and partners from the National Drought Mitigation Center, the National Center for Atmospheric Research (NCAR), and other groups. Their concerted research effort builds on individual MAPP research projects and related drought-research sector developments. The projects span the wide spectrum of drought research needed to make fundamental advances, from those aimed at the basic understanding of drought mechanisms to those evaluating new drought monitoring and prediction tools for operational and service purposes, and as part of NCEP's Climate Test Bed.

A major thrust of the DTF has been to develop a drought test-bed framework that individual research groups can use to develop and evaluate methods and ideas. Central to this is a focus on four high-profile North American droughts that are key areas for NIDIS early warning system development (1998-2004 western US drought, 2006-2007 SE US drought, the 2010- 2012 Tex-Mex drought over the Southern Plains, and the 2012 summer Midwestern US drought). The intent is to develop a framework that facilitates collaboration among projects, defines metrics to assess the quality of monitoring and prediction products, and helps to develop an experimental drought monitoring and prediction system that incorporates and assesses recent advances.

The DTF test-bed framework has been developed through the initial foundational efforts of three working groups (WG) that helped to address the major aspects of the test-bed: 1) WGI - Metrics: to define and apply metrics to evaluate advances in drought monitoring and prediction 2) WGII - Case Studies: to analyze drought cases by integrating all aspects of drought research and 3) WGIII - Experimental System: to incorporate research advances in an experimental drought monitoring and prediction system and assess improvements. More recently, a Research-to-Capability (RtC) activity has been initiated as part of the DTF with the goal of assessing recent progress in drought monitoring and prediction, with an eye towards advancing operational and service capabilities, building on the metrics and case studies framework developed by the WGs.

The DTF Assessment Protocol is presented as a reference for the groups that will engage in the DTF RtC activity, establishing guidelines for this assessment activity. The protocol may also be useful for drought researchers beyond the DTF effort. Scientists should be able to apply the common protocol to help provide quantitative answers to the basic question: ***Is my research effort improving upon current capabilities to monitor or predict drought, and by how much?***

1

http://cpo.noaa.gov/sites/cpo/Reports/MAPP/drought/DTF_Assessment_Protocol.pdf

UCAR

Drought Capability Assessment Protocol

Elements -- 1. Metrics

- ❑ draw from event verification practice (ie for floods)
- ❑ explicit inclusion of a benchmark capability or science
- ❑ require drought condition thresholds
- ❑ emphasize skill scores

Key predictand (s) for drought variable (e.g., P, T, soil moisture, streamflow)	**Metric(s) and skill scores comparing**
Onset and recovery of drought condition	Lead time of prediction Error of identification
Duration and severity of drought condition	Error, bias, correlation (time, value)
Indication (detection, prediction) of drought condition: deterministic	Categorical metrics: Critical Success Index (CSI), Equitable Threat Score (ETC) Probability of Detection (POD), False Alarm Rate (FAR), and others.
Probability of drought condition: probabilistic	Brier Skill Score (binary); secondarily, Brier decompositions for reliability and resolution
Value, overall Value given drought occurring in the observed or forecast period	1. Error, bias, correlation (of ensemble mean or median for probabilistic) 2. Ranked Probability Score (CRPS)

Drought Capability Assessment Protocol

Elements -- 2. Verification Datasets

Many verification data in drought categories and hydrologic fields are indices or ad hoc products. There is a need to be cautious on the uncertainties of all those products.

- **Precipitation**: surface rain gauge observations and blended precipitation analyses where appropriate (e.g., satellite, gage, radar blends of sufficient period coverage, extent and quality).

- **Temperature**: station observations and gridded analyses derived from station data and other sources, where appropriate.

- **Drought categories**: US Drought Monitor (USDM) categories may be used as verifying observations for categorical estimates or predictions unless other impact-based quantifications of drought existence or severity are available. In some cases it may be appropriate to verify categorical drought against univariate percentiles, e.g., from NLDAS soil moisture.

- **Hydrologic fields**: In-situ observations or derived analyses are a primary verification resource. ***Examples*** include soil moisture from NRCS SCAN or the North American Soil Moisture Data Base, snow water equivalent from SNOTEL or USHCN, snow cover from IMS, MODIS or Landsat, and streamflow from USGS gauge observations. For predictions, verification fields may also include observation-driven analyses or simulations (e.g., from NLDAS-2), or quality controlled input fields to the USDM. In general, verifying with monitoring simulations on other simulations is discouraged.

Drought Capability Assessment Protocol

Elements -- 3. Case Studies

- ❑ Forecast capability evaluation over a 30-year (1981-2010) period or longer is encouraged
- ❑ Attention to major drought case studies is also emphasized

The **four case studies** selected for drought capability evaluation are the following:

1) Winter 2001-Spring 2002 severe western US drought event.
   - Focus roughly on an area consisting of the 6-states CA, NV, UT, AZ, NM, and CO for Dec. 2001 thru May 2002, the primary wet season for most of that region
   - Evaluation of the overall 1998-2004 drought is also encouraged.
2) Fall 2005-Summer2008 sustained southeast US drought period
   - Focus roughly on an area consisting of the 4-states TN, MS, AL, GA, for which precipitation was mostly below average season-over-season beginning in Fall 2005 thru summer 2008. Rain began recovering in Fall 2008
3) The 2010- 2011 water-year drought over the Southern Plains,
   - Focus roughly on Texas, for the period beginning abruptly in Oct 2010 and continuing thru Sept 2011
4) The 2012 summer drought over the Central Great Plains,

Drought Capability Assessment Protocol

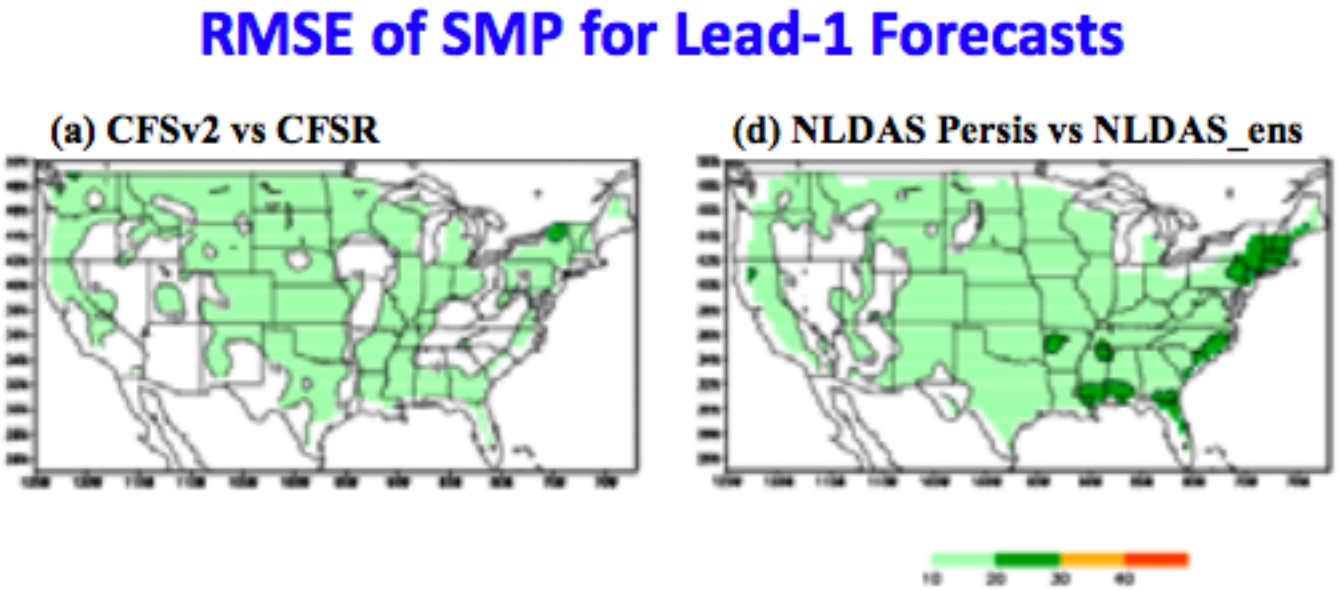
Elements -- 4. Baselines and Benchmarking

- ❑ Comparison to current capability baselines is critical
- ❑ A number of options are suggested in the Protocol document
- ❑ These depend on the application and will evolve…

The **baselines** against which research efforts are to be measured reflect existing operational or research capabilities. Primary baselines include:

- For monitoring or assessment capabilities
  - US Drought Monitor (USDM)
  - NLDAS Drought Monitor
  - SNOTEL-based analyses, e.g., SWSI
  - NCDC PDSI
  - VegDRI
- For prediction capabilities
  - CFSv2 or IRI's SPI forecast for atmospheric drought features (without further pre- or post- processing)
  - CPC Monthly and Seasonal Drought Outlooks
  - Streamflow predictions created via the Ensemble Streamflow Prediction (ESP) approach or by statistical water supply forecasting procedures (e.g., principle components regression), both of which represent current operational capabilities. Operational center datasets are preferred if available.
  - NCDC's PDSI forecasts, if appropriate

Example – Overall 'non-specific' metric



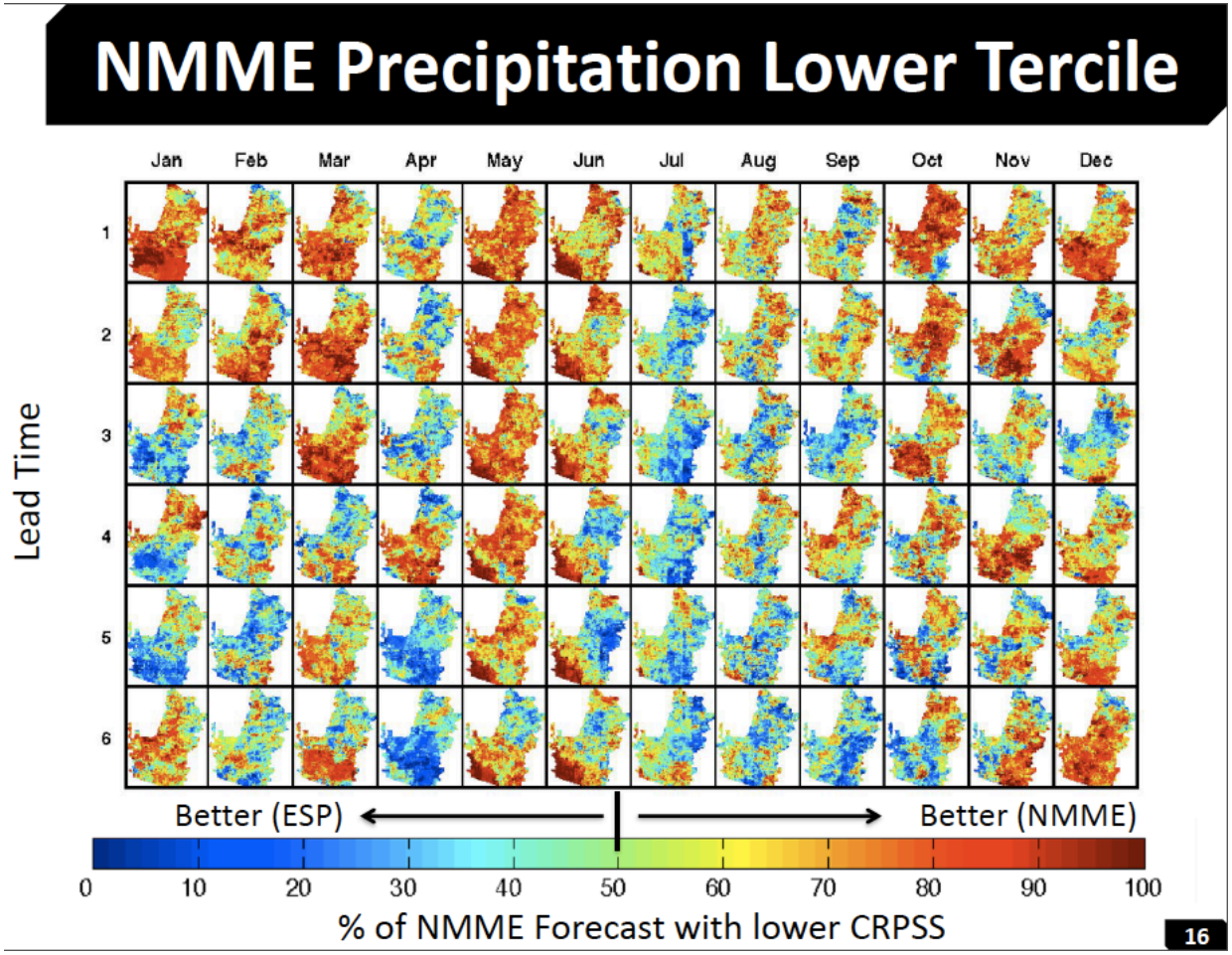
Chen & Mo (poster): **Seasonal Drought Prediction over the United States**

- a useful overall analysis – includes baselines – *but*:
- skewed/heteroskedasticvariable errors (runoff, snow, precipitation, flow) can be dominated by non-drought errors
- even for SM (perhaps beta distributed), non-drought behavior is likely the largest contribution to sample

NCAR
UCAR

Example – More drought-specific metric

- analyses focused on parts of the spectrum give insight
  - dry end – onset; wet end – recovery; seasonality important



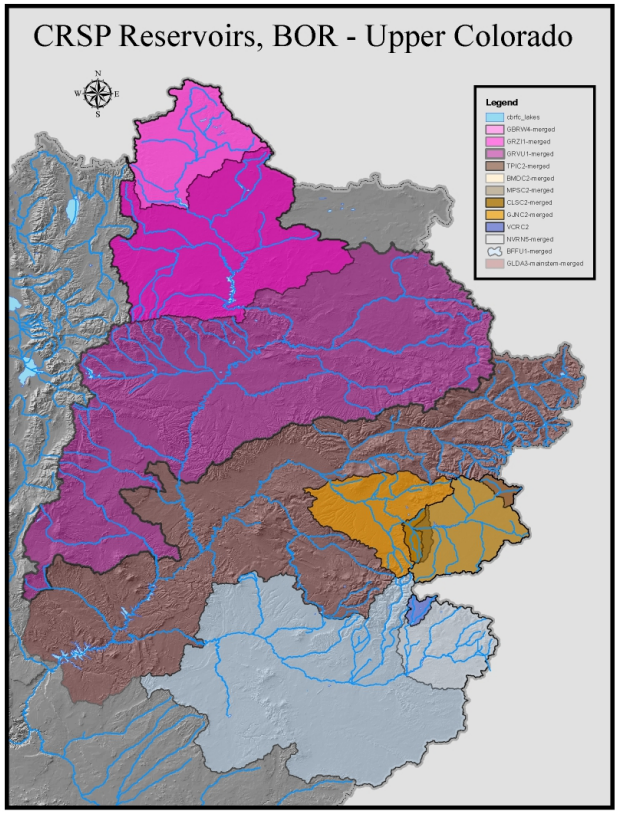
Wood (E) et al, CDPW 2012

NCAR UCAR

Application Example

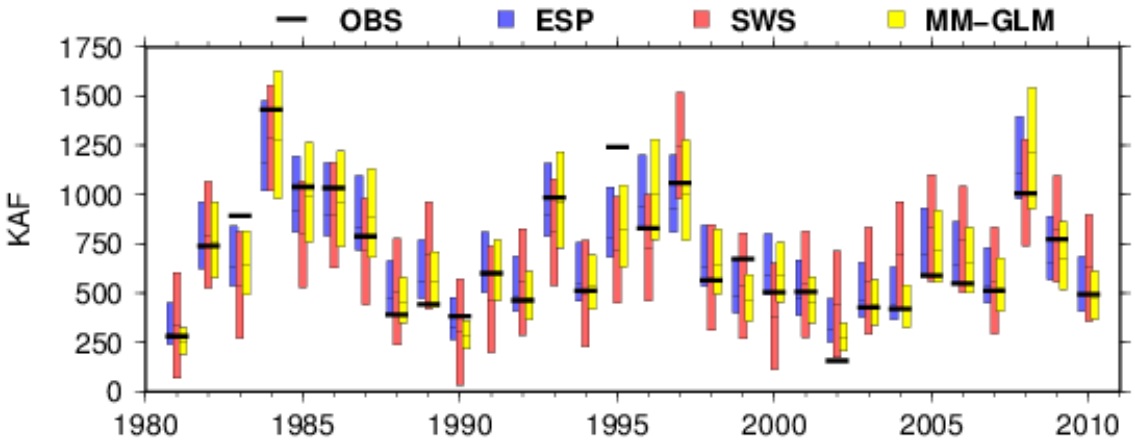
Western US Seasonal Water Supply Forecast (WSF)

- In Colorado R. basin, WSF (runoff volume predictions) can be a trigger for drought declaration, affecting 7 SWUS states
- NOAA/NWS makes 2 types of volume predictions in Jan – July period
  - ESP – model-based ensemble forecasts
  - SWS – statistical probabilistic predictions
- Subjectively merges them into a single 'NWS preferred' forecast



MAPP research question
Can these be objectively merged into a better WSF?

April 1 WSF for Gunnison R abv Blue Mesa Reservoir, CO



http://www.cbrfc.noaa.gov/testbeds/si_y2/

NCAR UCAR

Application Example

Multi-model Water Supply Forecast (WSF) research

- evaluated various approaches for merging ESP & SWS
  - *different GLMs, simple mean, Bayesian Model Avg (BMA)*
  - *focus here on 8 key locations for water supply management*

framework components	description
new capability	Objective Multi-model Water Supply Forecast
benchmark capability	**Baselines**: ESP, SWS, Simple Mean
verifying observation(s)	Observed naturalized runoff
analysis dataset(s)	Hindcasts of ESP and SWS for 30 years, 1981-2010 Cover 150-200 locations in the upper Colorado R. basin
supporting research	ample literature on WSF forecasting, less on streamflow forecast combination (eg Ajami et al)

key predictand(s) – for WSF	sample metric(s) – skill scores comparing:
value, overall	WSF median correlation by lead (deter.)
value indicating drought condition (probabilistic)	Brier skill score (binary) by lead

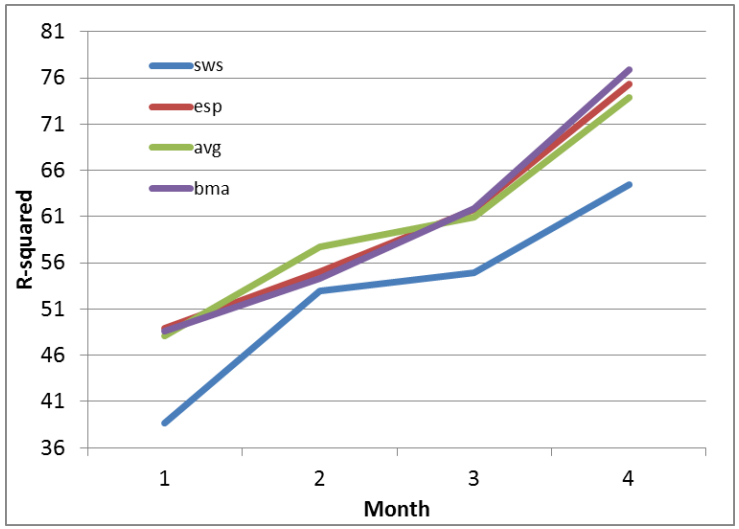
collaboration with James McCreigh and Balaji Rajagopalan (CU)

Application Example

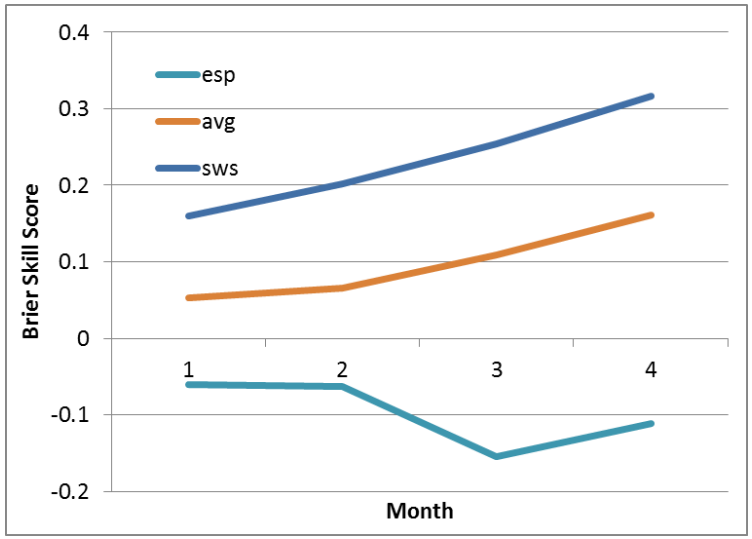
Western US Seasonal Water Supply Forecast (WSF)

- Pooling 8 key upper Colorado locations…
- On average, by these metrics, the BMA combination slightly improves upon the baseline forecasts in spring (March, April)
- From a *drought prediction* perspective, the BMA improves upon SWS and the simple mean, but *is worse than ESP in all months*

Overall Performance
Variance Explained



Drought Event Performance
Brier Skill Score relative to 3 baselines



NCAR UCAR

Summary and Future Directions

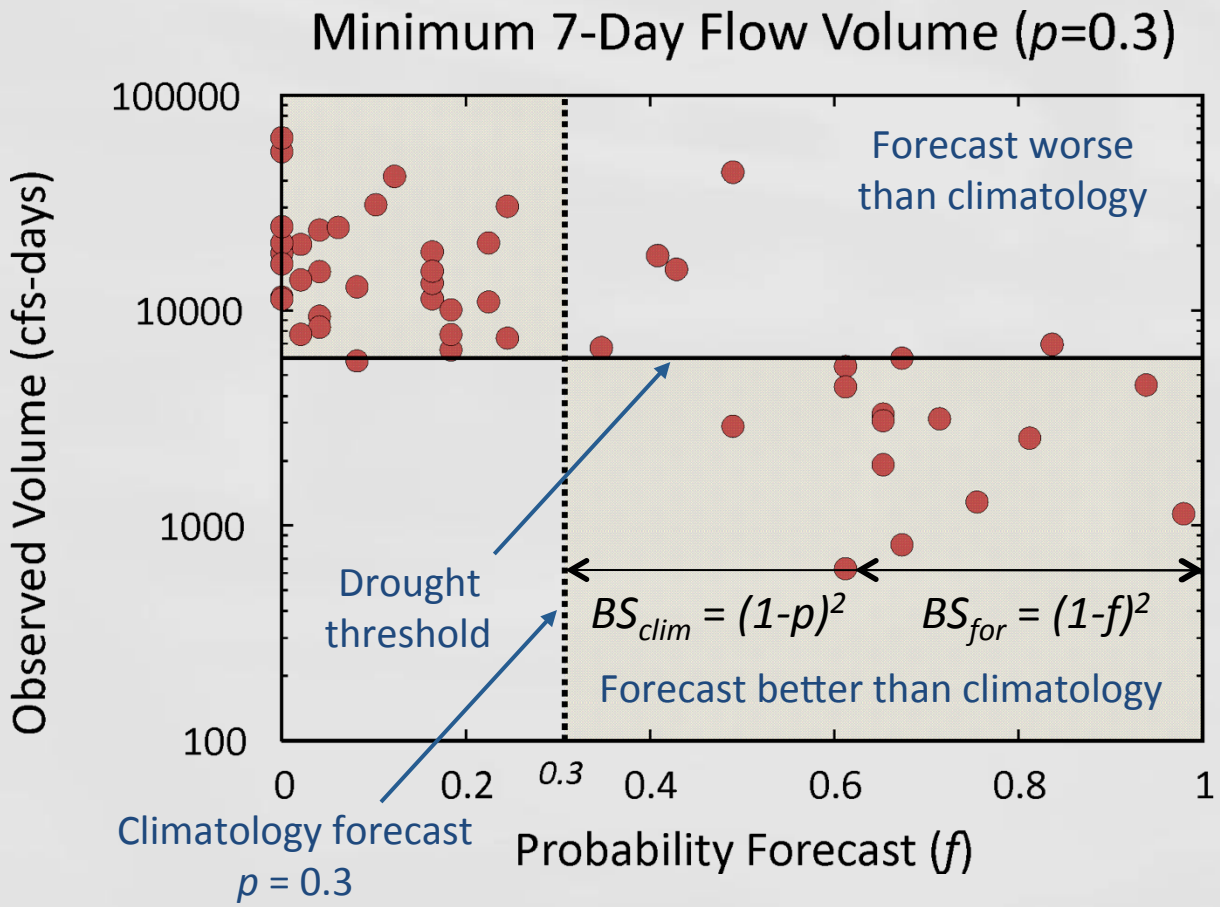
- The NOAA DTF is encouraging a more standard framework for evaluating drought prediction and monitoring capabilities and science
  - *phenomenological* – focus on drought states and features in addition to overall performance
  - measured against current benchmarks or operational baselines to show progress
  - recognizing drought science applications

- **Future Directions**
  - progressing on assessment using major US drought case studies (eg SE, Colorado, Tex-Mex/Midwest)
  - engaging partners in drought research nationally and internationally
  - ***encouraging drought researchers to use the Protocol to help quantify the benefit of their science advances for drought monitoring and prediction***


Questions?
andywood@ucar.edu

Drought Event Forecast Sample

All forecasts from a verification data set



Skill Score

$$SS = 1 - \frac{\overline{BS}_{for}}{\overline{BS}_{clim}}$$

$$= 0.578$$

Source: Allen Bradley

Applications – assessing our monitoring

Verification scores are straightforward for prediction… what about drought monitoring?

- There is a long standing concern about a lack of "observations" to verify 'drought monitoring'
  - e.g., can we accurately discriminate D1 from D3?

- The DTF focus is mostly on assessing our geophysical science, e.g.,
  - land surface modeling, climate prediction, climate simulation, remote sensing
  - such monitoring tools and data are verified against physical measurements in the traditional way
  - the framework goals apply

- For non-measured drought designations (eg D4), use of USDM as 'obs' has gained acceptance (note three previous talks)
  - it is a synthesis of many observational datasets, even if subjective
  - the Drought Impact Reporter (http://droughtreporter.unl.edu/) is also helpful for verification