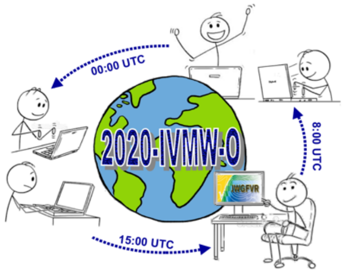


Program and abstracts of the November 2020 International Verification Method Workshop – Online

B. Casati (Editor)

Scientific committee:

Barbara Casati (MRD/ECCC, Canada)
Caio Coelho (CWFCS/NISR, Brazil)
Manfred Dorninger (University of Vienna, Austria)
Beth Ebert (Bureau of Meteorology, Australia)
Eric Gilleland (NCAR/RAL, USA)
Chiara Marsigli (DWD, Germany)
Marion Mittermaier (MetOffice, UK)

IT Support:

Markus Ristic (University of Vienna, Austria)
Pete Saddler (Shared Services Canada)

2020IVMWO sessions, program at a glance:

	00:00 UTC	8:00 UTC	15:00 UTC	20:00 UTC
04-Nov	IT test, ice-breaker	IT test, ice-breaker	IT test, ice-breaker	
09-Nov				OPENING + ERROR TRAKKING
10-Nov	DATA ASSIMILATION		REPRESENTATIVENESS + OBS.UNC.	
11-Nov			PROCESSES DIAGNOSTICS	HIW + USER VALUE CHAIN 20:00 UTC
12-Nov	HIW + USER VALUE CHAIN 00:00 UTC	HIW + USER VALUE CHAIN 8:00 UTC	PROCESSES+POLAR +VPROFILES	OCEAN
13-Nov			SEA-ICE	
16-Nov			SPATIAL METHODS 15:00 UTC	
17-Nov	SPATIAL METHODS 00:00 UTC	SPATIAL METHODS 8:00 UTC	METPLUS & OPS 15:00 UTC	
18-Nov	METPLUS & OPS 00:00 UTC	METAVERIF 8:00 UTC	METAVERIF 15:00 UTC	S2S,S2D,CLIMATE 20:00 UTC
19-Nov	METAVERIF 00:00 UTC		S2S,S2D,CLIMATE 15:00 UTC	
20-Nov				

Time-zone reference table for the 2020IVMWO sessions:

Time-zone reference table	**0:00-2:00 UTC**	**8:00-10:00 UTC**	**15:00-17:00 UTC**	**20:00-22:00 UTC**
Melbourne = UTC+11	11:00-13:00	19:00-21:00	2:00-4:00 (+1 day)	7:00-9:00 (+1 day)
Tokio = UTC+9	9:00-11:00	17:00-19:00	0:00-2:00 (+1 day)	5:00-7:00 (+1 day)
Beijing = UTC+8	8:00-10:00	16:00-18:00	23:00-1:00 (+1 day)	4:00-6:00 (+1 day)
New Delhi = UTC+5h30	5:30-7:30	13:30-15:30	20:30-22:30	1:30-3:30 (+1 day)
Moscow=UTC+3	3:00-6:00	11:00-13:00	18:00-20:00	23:00-1:00
Abuja = UTC+1	1:00-3:00	9:00-11:00	16:00-18:00	21:00-23:00
Paris = UTC+1	1:00-3:00	9:00-11:00	16:00-18:00	21:00-23:00
London = UTC	0:00-2:00	8:00-10:00	15:00-17:00	20:00-22:00
Buenos Aires = UTC-3	21:00-23:00 (-1 day)	5:00-7:00	12:00-14:00	17:00-19:00
Montreal = UTC-5	19:00-21:00 (-1 day)	3:00-5:00	10:00-12:00	15:00-17:00
Denver = UTC-7	17:00-19:00 (-1 day)	1:00-3:00	8:00-10:00	13:00-15:00
Vancouver = UTC-8	16:00-18:00 (-1 day)	0:00-2:00	7:00-9:00	12:00-14:00

Note: Paris = Barcellona = Bologna = Berlin = Oslo; Buenos Aires = Sao Paulo; grey shading indicates night-time.

Session program (1/7):

OPENING & ERROR TRACKING, 20:00-22:00 UTC, 9th Nov				
9 Nov, 20:00 UTC	Barbara Casati	Environment and Climate Change Canada (ECCC)	**A 20-year Journey of Forecast Verification Research**	**opening**
9 Nov, 20:40 UTC	Linus Magnusson	ECMWF, UK	**Understanding medium-range forecast errors from a synoptic-dynamic perspective**	**keynote**
9 Nov, 21:20 UTC	Shields Shannon	IMSG at NOAA NWS/NCEP/EMC, USA	**Analysis of Regional Sector Low-Skill Events in Recent Operational GFS Forecasts**	oral
DATA ASSIMILATION, 00:00-02:00 UTC, 10th Nov				
10 Nov, 00:00 UTC	Thomas Auligne	NCAR, USA	**Data Assimilation techniques in verification practices**	**keynote**
10 Nov, 00:40 UTC	Daisuke Hotta	MRI,JMA, Japan	**"Twin-analysis" verification: a new verification approach that alleviates pitfalls of "own-analysis" verification when applied to short-range forecasts**	oral
10 Nov, 01:00 UTC	Angela Cheng	McGill Univ, Canadian Ice Service (ECCC)	**Measuring uncertainty in sea ice concentration observations in Canadian Ice Service ice charts**	oral
REPRESENTATIVENESS & OBS. UNCERTAINTY, 15:00-17:00 UTC, 10th Nov				
10 Nov, 15:00 UTC	Barbara Casati	Environment and Climate Change Canada (ECCC)	**Representativeness issues in verification practices**	oral
10 Nov, 15:20 UTC	Nelson Shum	Environment and Climate Change Canada (ECCC)	**Representativeness of Coastal Stations for Verifying Open-Water 10 Metre Wind Forecasts**	oral
10 Nov, 15:40 UTC	Zied Ben Bouallegue	ECMWF, UK	**Accounting for representativeness in the verification of ensemble forecasts**	oral
10 Nov, 16:00 UTC	Morten Køltzow	MET Norway	**Taking the wind-induced undercatch of solid precipitation into account in the verification process**	oral
10 Nov, 16:20 UTC	Inoussa Abdou Saley	Faculty of S&T, Abdou Moumouni University of Niamey, Niger	**Are gridded datasets reliable for extreme rainfall events assessment over West African Sahel-Sudan?**	poster
PROCESSES DIAGNOSTICS, 15:00-17:30 UTC, 11th Nov				
11 Nov, 15:00 UTC	Thomas Haiden	ECMWF, UK	**Process-oriented verification**	**keynote**
11 Nov, 15:40 UTC	Mohana Thota	NCMRWF, India	**Representation of process based diagnostics in NCUM global and regional models**	oral
11 Nov, 16:00 UTC	Julian Quinting	Karlsruhe Institute of Technology, Germany	**Deep Learning for the Verification of Synoptic-scale Processes in NWP and Climate Models**	oral
11 Nov, 16:20 UTC	Jessica Baker	University of Leeds, UK	**Diagnosing land-atmosphere moisture coupling over South America in simulations from the UK and Brazil climate models**	oral
11 Nov, 16:40 UTC	Jason Otkin	University of Wisconsin-Madison, USA	**Evaluating the Impact of Planetary Boundary Layer, Land Surface Model, and Microphysics Parameterization Schemes on Cold Cloud Objects in Simulated GOES-16 Brightness Temperatures**	oral

Session program (2/7):

HIW & USER VALUE CHAIN, 20:00-22:00 UTC, 11th Nov (joint session with HIW webinars)				
11 Nov, 20:00 UTC	Beth Ebert	Bureau of Meteorology, Australia	**Overview of the HIWeather User-Oriented Evaluation Task Team Activities**	**keynote**
11 Nov, 20:40 UTC	Barbara Brown	NCAR, USA	**User-driven evaluation of tropical cyclone predictions**	oral
11 Nov, 21:00 UTC	Julia Chasco	Servicio Meteorológico Nacional, Argentina	**Incorporating the perspective of user evaluation into the creation of a new early warning system**	oral
11 Nov, 21:20 UTC	Amanda Anderson	UCAR, USA	**Verifying the Performance of a Coupled Fire-Atmosphere Model**	oral
HIW & USER VALUE CHAIN, 00:00-02:00 UTC, 12th Nov				
12 Nov, 00:00 UTC	Facundo San Martino	Servicio Meteorológico Nacional, Argentina	**Nowcasting verification in Argentina's Weather Service (SMN)**	oral
12 Nov, 00:20 UTC	David Wilke	Bureau of Meteorology, Australia	**Verification of a prototype wind impact forecast using building damage reports**	oral
12 Nov, 00:40 UTC	Jun Du	Environmental Modeling Center, NCEP/NWS/ NOAA, USA	**Measure of Forecast Challenge and Predictability Horizon Diagram Index for Ensemble Models**	oral
12 Nov, 01:00 UTC	Matt Boterhoven de Haan	Bureau of Meteorology, Australia	**Four Week Tropical Cyclone Forecast Verification**	poster
12 Nov, 01:10 UTC	Haoming Chen	Chinese Academy of Meteorological Sciences	**A verification method for sub-daily rainfall intensity in short-range forecast**	poster
HIW & USER VALUE CHAIN, 8:00-10:00 UTC, 12th Nov				
12 Nov, 8:00 UTC	Chiara Marsigli	Deutscher Wetterdienst, Germany	**Observations for high-impact weather and their use in verification**	**keynote**
12 Nov, 8:40 UTC	Mark Rodwell	ECMWF, UK	**User decisions, and how verification based the utility of these decisions could guide developments in probabilistic forecasting**	oral
12 Nov, 9:00 UTC	Michael Sharpe	UK Met Office	**New operational measure to assess extreme events using site-specific climatology**	oral
12 Nov, 9:20 UTC	Flora Gofa	Hellenic National Meteorological Service, Greece	**Appraisal of Challenging WeAther foREcasts (AWARE) in COSMO**	poster
12 Nov, 9:30 UTC	Stefano Materia	Fondazione CMCC, Italy	**Helping the agricultural food chain make better choices: The value of seasonal climate forecasts for European maize production**	poster
12 Nov, 9:40 UTC	Hellen Msemo	University of Leeds, United Kingdom	**Verification of Tanzanian Meteorological Authority Severe Weather impact-based forecasts**	poster

Session program (3/7):

PROCESSES+POLAR+VPROFILES, 15:00-17:00 UTC, 12th Nov				
12 Nov, 15:00 UTC	Day Jonathan	ECMWF, UK	**Measuring the impact of a new snow model using surface energy budget process relationships**	oral
12 Nov, 15:20 UTC	Solomon Amy	University of Colorado and NOAA/PSL	**Improving short-term forecasts of the ocean-sea ice-atmosphere coupled system using wintertime statistics from the MOSAiC campaign**	oral
12 Nov, 15:40 UTC	Borderies Mary	Meteo France	**The Most Resembling Column (MRC) validation method**	oral
12 Nov, 16:00 UTC	Petersen Raplh	University of Wisconsin-Madison, USA	**On the benefits of AMDAR Observation Profiles for Forecast Validation**	oral
OCEAN, 20:00-22:30 UTC, 12th Nov				
12 Nov, 20:00 UTC	Hernandez Fabrice	IRD/LEGOS (France) & UFPE/DOCEAN (Brasil)	**Measuring Performance, Skill and Accuracy in Operational Oceanography : Overview of approaches proposed by the GODAE/Ocean Predict Intercomparison and Validation Task Team**	**keynote**
12 Nov, 20:40 UTC	Mittermaier Marion	Met Office, UK	**Using MODE and MODE TD to investigate the evolution of the 2019 Chlorophyll-a bloom season in the North West European Shelf region**	oral
12 Nov, 21:00 UTC	Smith Gregory	Environment and Climate Change Canada	**Verification of eddy-properties in operational oceanographic analysis systems**	oral
12 Nov, 21:20 UTC	Aijaz Saima	Bureau of Meteorology, Australia	**Verification and inter-comparison of near-surface ocean currents in a global ocean forecasting system**	oral
12 Nov, 20:40 UTC	Clementi Emanuela	Centro Mediterraneo sui Cambiamenti Climatici, Italy	**The Mediterranean, Black and Marmara Seas analysis and forecasting physical systems: validation methodology and quality assessment**	poster
12 Nov, 20:50 UTC	Le Clainche Yvonnick	Fisheries and Oceans Canada	**Regional Class 4 verification of the Canadian operational ice-ocean prediction systems**	poster
SEA-ICE, 15:00-17:00 UTC, 13th Nov				
13 Nov, 15:00 UTC	Peterson Andrew	Environment and Climate Change Canada	**Using Integrated Ice Edge Error (IIEE) and Spatial Probability Score (SPS) to assess spread-error relationships in an ensemble sea ice forecast**	oral
13 Nov, 15:20 UTC	Zampieri Lorenzo	Alfred Wegner Institute, Germany	**Verification of subseasonal sea-ice prediction at both poles**	oral
13 Nov, 15:40 UTC	Niraula Bimochan	Alfred Wegner Institute, Germany	**Reference forecast of sea-ice edge using damped persistence of probability anomaly**	oral
13 Nov, 16:00 UTC	Cheng Angela	Environment and Climate Change Canada	**A symmetric spatial verification method for sea ice edges**	oral

Sessions program (4/7):

SPATIAL METHODS, 15:00-17:00 UTC, 16th Nov				
16 Nov 15:00 UTC	Barbara Brown	NCAR, USA	**MET and MesoVICT - Tools and Data for the Application and Testing of Established and New Spatial Verification Methods**	**joint keynote**
16 Nov 15:20 UTC	Manfred Dorninger	University of Vienna, Austria		
16 Nov 15:40 UTC	Gregor Skok	University of Ljubljana, Slovenia	**A new spatial displacement metric for continuous fields**	oral
16 Nov 16:00 UTC	Rachel North	Met Office, United Kingdom	**Using diagnostics from calculating verification scores to identify systematic errors**	oral
SPATIAL METHODS, 00:00-2:00 UTC, 17th Nov				
17 Nov, 00:00 UTC	Eric Gilleland	NCAR, USA	**Spatial Verification: A New Spatial Alignment Error Summary**	oral
17 Nov, 00:20 UTC	Patrick Skinner	CIMMS, NSSL, United States	**Object-based verification techniques for short-term thunderstorm forecasts**	oral
17 Nov, 00:40 UTC	Dominique Brunet	Environment and Climate Change Canada	**Decomposition of Verification Scores for Deterministic Continuous Forecasts: Insights from Image Quality Assessment**	oral
17 Nov, 01:00 UTC	Zhao Bin	NWP Centre, National Meteo. Centre, China	**Spatial verification for ensemble precipitation forecasts**	poster
17 Nov, 01:10 UTC	Natalí Aranda	National Meteorological Service, Argentina	**Spatial verification of high resolution precipitation forecasts over southern South America**	poster
SPATIAL METHODS, 8:00-10:00 UTC, 17th Nov				
17 Nov, 8:00 UTC	Bent Sass	Danish Meteorological Institute, Denmark	**Forecasting spatial structure of local precipitation extremes**	oral
17 Nov, 8:20 UTC	Seonaid R. Anderson	UK centre of ecology and hydrology	**The NFLICS project: Nowcasting FLood Impacts of Convective storms in the Sahel**	oral
17 Nov, 8:40 UTC	Joël Stein	Météo-France	**Neighborhood-based Continous Ranked Probability Score for Ensemble Prediction Systems**	oral
17 Nov, 9:00 UTC	Fabien Stoop	Météo-France	**Application of neighborhood-based contingency scores to AROME verification**	poster
17 Nov, 9:10 UTC	Carlo Cafaro	University of Reading, UK	**Do short-range convection-permitting ensembles lead to more skilful probabilistic rainfall forecasts over Tropical East Africa ?**	poster

Session program (5/7):

METPLUS and OPERATIONAL, 15:00-17:00 UTC, 17th Nov				
17 Nov, 15:00 UTC	Tara Jensen	National Center for Atmospheric Research, USA	**Fostering International Collaboration Through a Unified Verification, Validation, and Diagnostics Framework - METplus**	**keynote**
17 Nov, 15:40 UTC	Zhuo Wang	Illinois University, USA	**Process-oriented Model Diagnostics for Extended-range Forecasts**	oral
17 Nov, 16:00 UTC	Jonathan Vigh	National Center for Atmospheric Research, USA	**Developing a Space Weather Verification System Using METplus**	poster
17 Nov, 16:10 UTC	Perry Shafran	National Air Quality Forecasting Capability, NOAA/NWS, USA	**Verification of Air Quality Predictions Using METplus**	poster
17 Nov, 16:20 UTC	Jose Roberto Motta Garcia	CPTEC/INPE, Brazil	**MEC – A web-based tool for multi-model weather forecasting evaluation comparison**	poster
17 Nov, 16:30 UTC	Babatunde Atoyebi	Nigerian Meteorological Agency	**The dichotomous method of weather forecast verification at the central forecast office (CFO) of the Nigerian Meteorological Agency (NIMET)**	poster
METPLUS and OPERATIONAL, 00:00-02:00 UTC, 18th Nov				
18 Nov, 00:00 UTC	Jason Levit	Environmental Modeling Center, NCEP/ NWS/NOAA, United States	**Verification and Evaluation of Environmental Prediction Systems at the NOAA Environmental Modeling Center**	**keynote**
18 Nov, 00:40 UTC	Logan Dawson		**Methods and Tools Used to Verify Convection-Allowing Model Guidance at the NCEP/Environmental Modeling Center**	oral
18 Nov, 01:00 UTC	Yan Luo		**Exploring Spatial Distributions of Systematic Errors in the NCEP's Global Ensemble Precipitation Forecast Products**	oral
18 Nov, 01:20 UTC	Mallory Row		**A New METplus-based Verification System for the Global Forecast System (GFS)**	poster
18 Nov, 01:30 UTC	Alicia Bentley		**A New Webpage for Visualizing Verification Statistics from the National Centers for Environmental Prediction Production Suite**	poster
18 Nov, 01:40 UTC	Geoffrey Manikin		**The Model Evaluation Group at the Environmental Modeling Center**	poster

Session program (6/7):

METAVERIF, 8:00-10:00 UTC, 18th Nov				
18 Nov, 8:00 UTC	Keith Mitchell	Exeter University, UK	**Outcome-conditioned Decompositions of Proper Scores**	oral
18 Nov, 8:20 UTC	Martin Leutbecher	ECMWF, UK	**Understanding the link between ensemble mean error variance, spread-error ratio, mean error and the CRPS**	oral
18 Nov, 8:40 UTC	Alexander Jordan	Heidelberg Institute for Theoretical Studies, Germany	**Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited**	oral
18 Nov, 9:00 UTC	Sebastian Lerch	Karlsruhe Institute of Technology, Germany	**Evaluating probabilistic forecasts with scoringRules**	oral
METAVERIF, 15:00-17:00 UTC, 18th Nov				
18 Nov, 15:00 UTC	Harold Brooks	NOAA/National Severe Storms Laboratory	**The Relationship Between ROC, Performance, and the Quality-Decision Threshold Diagrams**	**keynote**
18 Nov, 15:40 UTC	Kenric Nelson	Photrek, USA	**Detecting over-confidence in weather forecasts**	oral
18 Nov, 16:00 UTC	Yawei Ning	Dalian University of Technology, China	**A new skill score for quantifying the uncertainty in multi-category precipitation forecasts**	oral
18 Nov, 16:20 UTC	Michael Sharpe	Met Office, UK	**A complementary measure to assess temporal uncertainty within Terminal Aerodrome Forecasts**	oral
METAVERIF, 00:00-02:00 UTC, 19th Nov				
19 Nov, 00:00 UTC	James Bennett	CSIRO, Australia	**How can we check the reliability of ensemble flood forecasts?**	**keynote**
19 Nov, 00:40 UTC	Nachiketa Acharya	International Research Institute for Climate and Society (IRI), Columbia University, USA	**Point-Biserial Correlation-Based Skill Scores for Probabilistic Forecasts**	oral
19 Nov, 01:00 UTC	Deryn Griffiths	Bureau of Meteorology, Australia	**Verification of Quantile Forecasts – A Journey**	oral
19 Nov, 01:20 UTC	Rob Taggart	Bureau of Meteorology, Australia	**Huber loss as a scoring function for forecast verification**	oral

Session program (7/7)

S2S,S2D,CLIMATE, 20:00-22:00 UTC, 18th Nov				
18 Nov, 20:00 UTC	Francisco Doblas-Reyes	Barcelona Supercomputing Center, Spain	**Forecast quality assessment for operational climate prediction**	**keynote**
18 Nov, 20:40 UTC	Andrea Manrique-Suñén	Barcelona Supercomputing Center, Spain	**Choices in the verification of S2S forecasts**	oral
18 Nov, 21:00 UTC	Caio Coelho	CPTEC/INPE, Brazil	**Evaluating the representation of precipitation variability patterns over South America in sub-seasonal predictions**	oral
18 Nov, 21:20 UTC	Felipe Andrade	University of Reading, UK	**Evaluation of sub-seasonal precipitation forecasts for Africa**	oral
18 Nov, 21:40 UTC	Ángel G. Muñoz	International Research Institute for Climate and Society (IRI). Columbia University. USA	**Sub-Seasonal Forecast Skill: When, Where and How To Find It?**	oral
S2S,S2D,CLIMATE, 15:00-17:00 UTC, 19th Nov				
19 Nov, 15:00 UTC	Elizabeth Weatherhead	Colorado University, Boulder, CO, USA	**Validation and Verification of Climate Products**	**keynote**
19 Nov, 15:40 UTC	Joshua French	University of Colorado, Denver, USA	**Nonparametric Permutation Procedures for Evaluating Climate Model Accuracy**	oral
19 Nov, 16:00 UTC	Nathan Lenssen	IRI, Columbia University, US	**Seasonal Forecast Skill of ENSO Teleconnection Maps**	oral
19 Nov, 16:20 UTC	Dominik Büeler	Karlsruhe Institute of Technology, Germany	**The role of model calibration and verification techniques for sub-seasonal weather regime forecast skill**	oral
19 Nov, 16:40 UTC	Andreas Paxian	Deutscher Wetterdienst, Germany	**User-oriented verification of the DWD climate prediction website**	oral

Abstracts are presented in alphabetical order,
according to the family name of the presenting author.

Title: Are gridded datasets reliable for extreme rainfall events assessment over West African Sahel-Sudan?

Authors: Inoussa Abdou Saley[1] and Seyni Salack
1 = Faculty of Sciences and Technology, Abdou Moumouni University of Niamey, Niger

In-situ gauge records are used to measure the amount of rainfall fallen at a given point. They remain the most accurate information for monitoring location-specific extremes. However, inaccessibility to some remote areas, the limited spatial coverage of raingauges, and the advances of earth observing systems are favoring an increased use of gridded precipitation datasets to deliver climate information services over the West African Sahel (WAS).

Are gridded datasets reliable for monitoring heavy rain events over this region? To answer this question, daily precipitation from twelve (12) of the most commonly used gridded datasets are investigated including satellite estimates, interpolated rain-gauge data, reanalysis, and merged products. They were assessed, against a quality controlled in-situ data from sixty-nine (69) stations of WAS countries, using an unsupervised clustering algorithm which led to the identification of three (03) categories of heavy rain events. At an explained variance exceeding 60%, all gridded datasets exhibited 03 categories of heavy rains but of lower intensities and overestimated frequency of occurrences. ARC2 satellite showed the best intensity-duration-frequency balance but trends are better depicted by GPCC and TAMSAT products. Even though, in-situ gauge records are preferred for monitoring extreme rainfall, the use of gridded precipitation products need account for the relative definition of heavy rain events in the involved datasets.

Title: Point-Biserial Correlation-Based Skill Scores for Probabilistic Forecasts

Authors: Nachiketa Acharya[1] and Michael K. Tippett

1 = International Research Institute for Climate and Society (IRI), Columbia University, USA

The point-biserial correlation (rpb) coefficient is a measure of the strength of association between a continuous-level variable and a dichotomous ("naturally" or "artificially" dichotomized) variable. The rpb is mathematically equivalent to Pearson correlation but has a more intuitive formula which provides insights on what constitutes a "good" association between continuous and dichotomous variable. In the probabilistic forecasts verification system, skill scores are estimated between issued forecast probabilities (continuous variable) and relative observed category (whether or not the event; dichotomous variable). Most of the existing skill scores for probabilistic forecasts focusing either on the mean squared error in probabilistic space (Brier score) or degree of correspondence between issued forecast probabilities and relative observed frequencies (reliability diagrams) or the degree of correct probabilistic discrimination in a set of forecasts. In this study, we will introduce the use of rpb to verify probabilistic forecasts for measuring the strength of association between issued forecast probabilities and actual observed events. The proposed method will be demonstrated in experimental evaluation with synthetic and real precipitation forecasts.

Title: Verification and inter-comparison of near-surface ocean currents in a global ocean forecasting system

Authors: Saima Aijaz[1], Gary Brassington, Prasanth Divakaran, Charly Regnier
1 = Bureau of Meteorology, Australia

The OceanPredict task team for Intercomparison and Verification (IV-TT) has established the CLASS4 data standard for routine verification against reference observing platforms. The near-surface currents derived from the trajectories of drifting buoys drogued at 15m have recently been added to the CLASS4 reference data. The drifter buoy dataset is managed and operated by the Coriolis Datacenter (Ifremer and Météo-France), and is accessed by several national centres. We have recently applied this data to the Ocean Model, Analysis and Prediction System (OceanMAPS) at the Australian Bureau of Meteorology. Our motivations for this verification include: 1) Inform our stakeholders of the performance; 2) extend the routine monitoring of the operational system; and 3) inform the future research and development. We verify the 24-hour average best estimates and forecasts of currents extracted at 15m against the observations from 2018 and 2019. We compute statistical metrics for two separate regions: global and Australian (0-50S, 90-180E). We further compare OceanMAPS currents with those from the Copernicus Marine Environment Monitoring Service (CMEMS) model (NEMO) developed by Mercator Océan. Our results show that the OceanMAPS has good skill in reproducing the 24-hour average near-surface currents. The average annual biases in zonal and meridional velocities are less than 0.04 m/s and RMS error ranges from 0.2 m/s to 0.24 m/s. The performance of the global region is marginally better than the Australian region. The near-surface currents predicted by OceanMAPS and CMEMS (NEMO) are in excellent agreement despite the systems having independent ocean models and data assimilation methods.

Title: Verifying the Performance of a Coupled Fire-Atmosphere Model

Authors: Amanda R. Siems Anderson
National Center for Atmospheric Research, USA

Fire spread models are a useful decision support tool for agencies responding to wildfires, aiding them in allocating their resources both in quantity and in spatial distribution by predicting fire behavior. The Colorado Fire Prediction System (CO-FPS) was developed by the National Center for Atmospheric Research (NCAR) in collaboration with the Colorado Division of Fire Prevention and Control's Center of Excellence for Advanced Aerial Firefighting to provide such decision support to fire response managers in Colorado, USA. This coupled atmospheric-fire spread model allows personnel to run simulations via a web interface to receive model forecasts of fire behavior including spread, flame length, and smoke release.

However, a successful model deployment requires confidence in the model's ability to accurately predict fire behavior. While the atmospheric component of the model can rely on well-developed verification techniques that take advantage of the wide number of atmospheric observations available across the state of Colorado and surrounding areas, evaluating the fire behavior portion of the model presents a larger challenge. Fire-related observations are often collected during fire response operations specifically to be used by responders in the moment – this means they may lack accuracy, timeliness, or relevance, and they are often stored in disparate databases, in physical copies in filing cabinets, or not at all.

This presentation will focus on the verification methods and observations sets used during the development of CO-FPS. These included traditional methods adapted to use for fire spread, such as contingency statistics typically used for precipitation accumulation, along with personal interactions with CO-FPS end-users. Difficulties with the availability and quantity of observations will also be discussed, along with using observations of opportunity including social media posts and conversations with responders.

Title: The NFLICS project: Nowcasting FLood Impacts of Convective storms in the Sahel

Authors: Seonaid R. Anderson[1], Steven J. Cole, Cheikh Abdoulahat Diop, Christopher M. Taylor, Cornelia Klein
1 = UK Centre of Ecology and Hydrology

Flash flooding from intense rainfall frequently results in major damage and loss of life across Africa. In the Sahel, intense rainfall from Mesoscale Convective Systems (MCSs) is the main driver of flash floods, with recent research showing that these have tripled in frequency over the last 35 years. The project NFLICS (Nowcasting FLood Impacts of Convective storms in the Sahel) is developing a prototype early warning system for Senegal, nowcasting convective activity and flood risk from MCSs at city and sub-national scales out to 6-hours.

Nowcasts are based on wavelet analysis of historical and real-time Meteosat data on cloud-top temperature, conditioned on the present location and timing of observed convection. Verification against the forecast analysis assessed the utility of the probabilistic nowcasting method for predicting future convective structures. As spatial uncertainties were found to be a key influence on performance, a neighbourhood approach was applied, taking the probability of convective activity occurring anywhere within a neighbourhood surrounding each grid-point. By verifying the nowcast skill against a climatological reference, the neighbourhood size leading to the highest nowcast skill was identified at each lead-time. Verification of the maximum forecast probability of convection, against 24-hour raingauge accumulations over Dakar, allowed assessment of the probabilistic nowcast skill for forecasting extreme precipitation events.

Although this verification is over a limited geographical area, and the definitions of forecast and observed events differ, this assessment against independent observations is crucial to understand the nowcast performance for predicting extreme flood-producing events in an operational context.

Title: Evaluation of sub-seasonal precipitation forecasts for Africa

Authors: Felipe M. de Andrade[1], Matthew P. Young, David MacLeod, Linda C. Hirons, Steven J. Woolnough and Emily Black
1 = University of Reading, UK

Sub-seasonal precipitation forecasts for Africa are evaluated using hindcasts from three operational models (ECMWF, UKMO and NCEP) participating in the Subseasonal to Seasonal (S2S) prediction project. A variety of verification metrics are employed to assess the quality of both deterministic and probabilistic forecasts of weekly accumulated precipitation at lead times of one to four weeks ahead during different rainy seasons. The metrics included mean error, Pearson's correlation, mean square skill score, discrete ranked probability skill score, relative operating characteristic and attributes diagram. Models have better performance in the first two weeks of forecasts compared to subsequent weeks. Deterministic and probabilistic forecast quality assessments indicate reasonable agreement among different verification metrics analysed, showing overall more skilful predictions for ECMWF over East Africa compared to other models and regions. Tercile-based probabilistic forecasts reveal roughly similar characteristics between outer categories and low quality on near-normal category. Overconfident forecasts are verified for all weeks and models, needing to apply calibration techniques for providing more reliable weekly precipitation predictions for Africa.

Title: Spatial verification of high resolution precipitation forecasts over southern South America

Authors: Aranda, Natalí Giselle[1]; García Skabar, Yanina; Matsudo, Cynthia Mariana
1 = National Meteorological Service, Argentina

The method for object-based diagnostic evaluation (MODE) is a spatial verification method that attempts to identify regions of interest, like precipitation, in the same way that a human would do. This method defines objects in the forecast and observation fields based on user-defined parameters. MODE was used to evaluate the performance of 4-km hourly precipitation forecasts from the Weather and Research Forecasting Model (WRF) over southern South America against GPM derived product IMERG Final Run version. For February 2018, multiple tests were performed for the threshold and the radius of convolution parameters, to select adequate values for 3 and 24 h accumulated precipitation. To detect errors in synoptic and convective-scale systems, the study showed it could be used a smoothing radius of 50 km, a threshold of 3 mm for 3-h rainfall accumulation and 10 mm in a period of accumulation of 24 h. Furthermore, precipitation forecasts verification were made for 2017-2018 two-year period. The 3-h analysis exposed the spin up time in the first three hours. Also, it showed the relationship between forecasted and observed objects remained high during the 48 hours. The daily precipitation study showed the performance of the model improved during the winter. Moreover, in both evaluations, MODE and traditional verification statistics (eg, Probability of Detection, False Alarm Ratio) were used as complementary verification methods. Traditional verification allowed to understand the overall performance of the model, including false-alarms and misses, and MODE proved the WRF had low errors associated with the location, coverage area and intensity.

Title: The dichotomous method of weather forecast verification at the central forecast office (CFO) of the Nigerian Meteorological Agency (NIMET).

Authors: Babatunde Atoyebi[1], Ugochukwu Ezedigboh, Linda Nwachukwu, Stella Afolayan, Abayomi Abiola Okanlawon, Desmond Onyilo
1 = Nigerian Meteorological Agency, Abuja, Nigeria

The dichotomous method is a method that works simply and effectively well for the team of meteorologists, who carry out both daily validations and monthly verifications of weather forecast issued for 48 major cities spread across the length and breadth of the 36 states of Nigeria, at the CFO NIMET Agency in Abuja, Nigeria. At first, separate validation and verification are carried out for “am” periods and “pm” periods, representing times that span from 12 midnight to 11:59am in the morning and from 12 mid day to 11:59pm in the night respectively. After which the average for the day is computed. We use “YES” and “NO” to represent an event and a non-event respectively. A total of four(4) cells for every station (to represent; morning observation, afternoon observation, morning forecast and afternoon forecast) are used to determine hits, misses, correct non-events and false alarms, from which the Accuracy, Bias, Probability of Detection, False Alarm Ratio, Probability of False Detection and Critical Success Index for the weather forecasts are calculated for each month. Some Microsoft Excel's logical, statistical and arithmetical functions have really been instrumental in achieving this task of weather forecast verification at the CFO of NIMET Agency. Below is the outcome of weather forecast verification for July 2020 by using this dichotomous method. MONTHLY AVERAGE: Accuracy 0.759, Bias 1.407, Probability of Detection 0.632, False Alarm Ratio 0.432, Probability of False Detection 0.190, Critical Success Index 0.422.

Title: Diagnosing land-atmosphere moisture coupling over South America in simulations from the UK and Brazil climate models

Authors: Jessica C.A. Baker[1], Dayana Castilho de Souza, Paulo Kubota, Wolfgang Buermann, Caio A.S. Coelho, Martin B. Andrews, Manuel Gloor, Luis Garcia-Carreras, Silvio N. Figueroa and Dominick V. Spracklen

1 = University of Leeds, UK

Assessing model representation of land-atmosphere interactions is essential to evaluating model performance, though evaluation of these processes in climate models has so far been limited. We developed a new diagnostic toolkit to assess land-atmosphere moisture coupling in satellite observations, reanalysis data and simulations from two global climate models: the Brazilian Global Atmospheric Model version 1.2 (BAM-1.2) and the UK Hadley Centre Global Environment Model version 3 (HadGEM3). Our analysis focussed on South America where land-atmosphere interactions have an important impact on climate, though the diagnostics illustrated here can be applied to any region of interest. Multiple metrics were used to measure the strength of coupling between the land and the atmosphere, identify whether surface moisture fluxes are controlled via a land-surface or atmospheric forcing, and examine how controls vary spatially and throughout the seasonal cycle. Feedback pathways are traced from the surface to the atmosphere in a mechanistic way to fully understand model representation of physical processes and identify targets for model development. Both models capture key features of South American land-atmosphere moisture coupling, including seasonal variation in coupling strength, large-scale spatial variation in the sensitivity of evapotranspiration to soil moisture, and a northeast-southwest dipole in evaporative regime across the continent. However, weaknesses are also identified, with HadGEM3 and BAM-1.2 sometimes misrepresenting the strength or direction of interactions, for example over parts of Amazonia. Where these models are unable to accurately simulate land-atmosphere moisture feedbacks, precipitation biases and misrepresentation of processes controlling soil moisture are implicated as likely drivers.

Title: Accounting for representativeness in the verification of ensemble forecasts

Author: Zied Ben Bouallegue[1]
1 = European Centre for Medium-Range Weather Forecast, UK

If observation uncertainty is not accounted for when verifying ensemble forecasts, then the investigator may draw inappropriate conclusions about the performance of the prediction system. In order to account for observation uncertainty in the ensemble verification, observation errors have first to be characterized. Representativeness errors are assumed to be the predominant contribution to observation errors associated with station measurements in our applications. So, the question is to what extent a measurement at a single location is representative of the situation over a larger area. Characterization of representativeness error is made in probabilistic terms using a parametric approach, that is by fitting a probability distribution. The shape of the distribution depends on the weather variable of interest while its parameters are estimated with the help of a high-density network of stations over Europe. Based on this analysis, uncertainty associated with the scale mismatch between forecast and observation can be accounted for in the verification by applying the so-called perturbed ensemble approach. Verification results show a large impact of representativeness on forecast reliability and skill estimates.

Title: How can we check the reliability of ensemble flood forecasts?

Authors: James Bennett[1] & David Robertson
1 = CSIRO, Australia

Flood prediction is a crucial task for ensemble streamflow forecasts. For ensemble flood forecasts to be effective, the probabilities drawn from the ensemble – e.g. the probability of exceeding a flood threshold – must be meaningful. That is, ensemble flood forecasts must be statistically reliable. Despite this, reliability of ensemble flood forecasts is rarely checked. This is in part because of the practice of verifying retrospective flood forecasts mainly on large historical floods. Selecting only observed floods for verification precludes a formal assessment of reliability. This problem can be avoided by selecting forecasts for verification based on whether the forecast indicates a heightened probability of a flood, irrespective of whether a flood was then observed. We present a method for selecting forecasts that allows the assessment of reliability. Once forecasts are selected, we demonstrate a new method for assessing the reliability of forecasts of peak streamflow, based on the widely used probability integral transform. We show that these methods are more useful for measuring the performance of ensemble flood prediction system than conventional methods that focus on performance of forecasts only for historical floods.

Title: A New Webpage for Visualizing Verification Statistics from the National Centers for Environmental Prediction Production Suite

Author: Alicia M. Bentley[1], Mallory P. Row, Logan C. Dawson, and Jason J. Levit
1 = NOAA/NCEP/EMC, USA

The Verification, Post-Processing, and Product Generation (VPPPG) branch of the National Centers for Environmental Prediction (NCEP)/Environmental Modeling Center (EMC) has developed a new verification webpage to consolidate and disseminate verification statistics and graphics associated with models that are included in the NCEP Production Suite. The development of this new verification webpage coincides with NCEP/EMC's transition from producing verification statistics and graphics using the Verification Statistics Database (VSDB) developed internally to using the Model Evaluation Tools (MET) and METplus developed externally at the National Center for Atmospheric Research (NCAR)/Developmental Testbed Center (DTC). In addition to displaying operational verification statistics and graphics produced using MET and METplus, the new verification webpage provides the framework for allowing users to access experimental model verification statistics and graphics during official model evaluations. This presentation provides a first look at 1) the design of the new NCEP/EMC verification homepage, 2) the ability of individual model verification webpages to display operational and experimental verification statistics and graphics, as well as 3) the new NCEP/EMC verification graphics generated using METplus and its suite of Python wrappers.

Title: Spatial verification for ensemble precipitation forecasts

Author: Zhao Bin[1]
1 = Numerical Weather Prediction Centre, National Meteorological Centre, China

With continuous improvement in computer power and data simulation techniques, higher-resolution ensemble systems have been available to meet the forecast requirements of small-scale local weather events, especially extreme weather such as heavy rainfall and tropical cyclones. The spatial methods are mainly proposed for deterministic simulations and rare research has been addressed the spatial verification of ensemble forecasts. While the ensemble forecast model has also developed to higher resolution. It is well known that the ensemble forecast can fully resolve the uncertainties of small-scale processes. However, the small-scale spatial and temporal mismatching is still adheres to high resolution ensemble forecast. The “double penalty” in high resolution due to the initial and model perturbation is not reduced even if the ensemble members are increased because of the limitation on the probability field. Therefore, we need to apply the spatial verification methods to the ensemble precipitation verification. In this paper, based on CRPS and upscaling method just like FSS, a more proper and reasonable spatial verification method for ensemble quantitative precipitation forecast has been designed and proposed and the assessment of advantage is explored with the thresholds as well as through a direct comparison with traditional ensemble skill score and certain simple upscaling verification technique as reference.

Title: The Most Resembling Column (MRC) validation method

Authors: M. Borderies[1], O. Caumont, C. Augros, E. Bresson, J. Delanoë, V. Ducrocq, N. Fourrié, T. Le Bastard and M. Nuret
1 = Meteo France

Validating the representation of clouds and precipitation in atmospheric models faces many difficulties. Classical metrics based on point-to-point comparisons suffer from the well-known "double penalty" effect. Conversely, metrics that ignore the geographical position of the observation and forecast do not take sufficient advantage of the information conveyed by the observations. While there are intermediate neighbourhood or object-based methods, these are often two-dimensional in the horizontal plane. When they are three-dimensional, they also require three-dimensional observations in order to match structures or objects that are both present in the observations and the forecasts. To circumvent these problems, we present here a validation method called the Most Resembling Column (MRC, Borderies et al. 2018) that is particularly well suited for observations in the form of vertical profiles such as those observed by cloud radars and lidars, but which is also applicable to volumes of observations. This type of validation makes it possible to separate positioning errors from other errors in the model or observation simulator (e.g. microphysical properties).

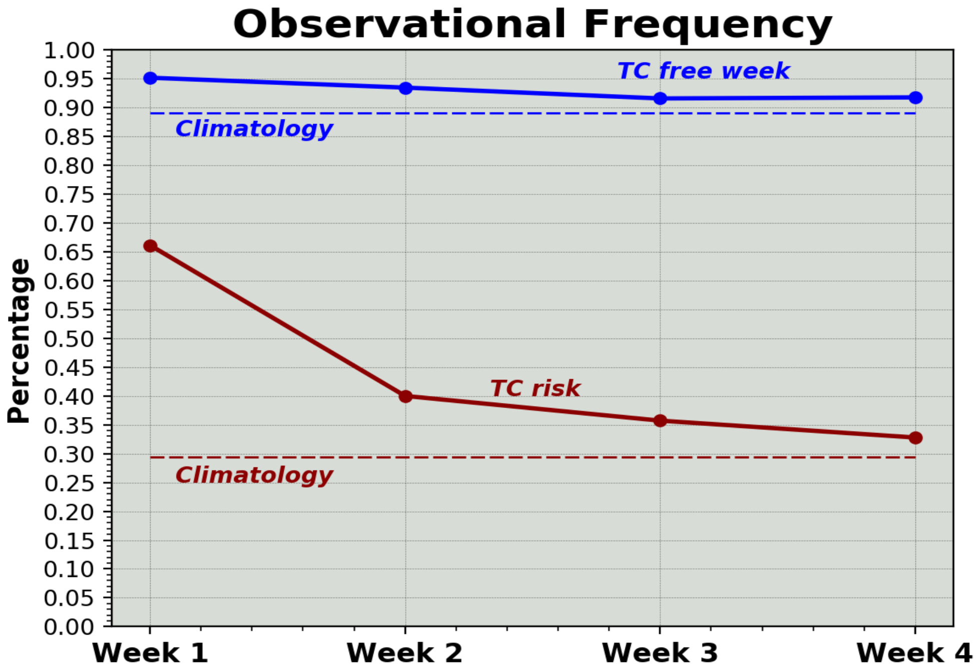
After describing the MRC method, the W-band reflectivity observation operator designed for regional convective-scale numerical weather prediction (NWP) systems like AROME-WMed (Fourrié et al. 2015) is validated with data from the airborne cloud radar RASTA (Delanoë et al. 2013) over a 2-month period. Finally, the MRC method is used to calibrate the W-band radar observation operator by retrieving the optimal effective shapes of the predicted graupel, snow and pristine ice, through minimization of the standard deviation between observations and simulations.

Title: Four Week Tropical Cyclone Forecast Verification

Author: Matthew A. Boterhoven de Haan[1]
[1]= Bureau of Meteorology, Australia

Tropical cyclones are among the most destructive natural phenomena. Preparing for a tropical cyclone impact is essential in tropical cyclone prone regions, especially for the oil and gas industry that require multiple days of preparedness. There is also a need to conduct maintenance activities and there is a desire to know about tropical cyclone free weeks in advance so the maintenance can be planned. Tropical cyclone meteorologists issue weekly tropical cyclone categorical ratings (Very Low (<5%), Low (5-20%), Moderate (20-50%), and High(>50%)) in the four week tropical cyclone forecast for the Northwestern region of Australia (105E-130E).

To produce verification from the categorical ratings we assume the customer is interested in identifying risk and non-risk periods. For risk periods tropical cyclones were observed around 66% (week 1) to 33% (week 4) of the time. For all weeks the forecast of a risk period was better than climatology. For the non-risk periods the forecast improved on the climatological forecast for all four weeks. The verification results demonstrate that the Four Week Tropical Cyclone Forecast has a degree of skill versus climatology.



Title: The Relationship Between ROC, Performance, and the Quality-Decision Threshold Diagrams

Authors: Harold E. Brooks[1] and James Correia Jr.
1= NOAA National Severe Storms Laboratory, USA

Forecasts of dichotomous events under uncertainty with varying weights of evidence can be evaluated by setting varying thresholds across the weight of evidence creating a series of 2x2 problems. Multiple aspects of forecast performance can be visualized on diagrams such as the relative operating characteristics and performance diagrams. Using a simple model of the problem based on signal detection theory using Gaussian distributions for the events and non-events, we have investigated the display of performance on the different diagrams as a function of the base rate of the event, the separation of the events and non-events, and the difference in the variance of the distributions. We have developed a new diagram, the Quality-Decision Threshold diagram, which explicitly separates the ability of a system to discriminate between event and non-events, and the decision threshold. After considering the idealized framework, we show examples from a series of forecasts related to severe thunderstorms and tornadoes.

Title: User-driven evaluation of tropical cyclone predictions

Author: Barbara Brown[1], Louisa Nance[1], Christopher Williams[2]
1 = National Center for Atmospheric Research, Boulder, CO, USA
2 = University of Florida Department of Geography, Gainesville, FL, USA

The evolution of verification approaches in recent years has included a focus on development and application of user-relevant metrics for evaluation of weather and climate forecasts. The goal of such efforts is to provide information that is meaningful for specific forecast users. User-relevant verification methods avoid the too-common approach of applying the same standard metric for all applications, and instead focus on questions of interest to particular users. The process of creating such metrics involves clearly identifying the questions of interest to forecast users and then defining metrics that can answer those questions. This presentation will briefly consider user-relevant verification concepts in general, with a specific example provided by recent studies associated with the US Hurricane Forecast Improvement Project (HFIP), focused on evaluations of tropical cyclone (TC) predictions.

Each year, TCs cause significant property damage and human impacts (death, injury, displacement) around the world. To mitigate these impacts, weather prediction centers produce forecasts of TC movement (i.e., track) and intensity, with warnings based on these forecasts provided to emergency managers and the public. Guidance used by operational forecasters includes predictions of storm motion and intensity from operational numerical weather prediction (NWP) models. In response to needs for improved predictions of TC track and intensity (with a major focus on intensity), the US National Weather Service implemented the Hurricane Forecast Improvement Project (HFIP) in 2007, with a goal of making significant improvements to TC track and intensity predictions. HFIP engaged NWP model developers to improve the TC guidance produced their models. In each of five years, scientists at the National Center for Atmospheric Research evaluated operational and experimental models to help select new modeling systems to demonstrate to forecasters at the US National Hurricane Center (NHC) during the subsequent hurricane season.

Each year, the capabilities of each experimental model were compared to the performance of predictions from current “baseline” operational models. Questions to be addressed were elicited from NHC forecasters and managers. Once the questions were identified, appropriate statistically-valid verification methods were developed to evaluate the predictions from the various experimental models in comparison to the predictions from the baseline models. This process resulted in meaningful, actionable, verification information, used to select the demonstration models each year.

This presentation will describe the process of developing the verification approaches and the rationale behind each method that was applied. In addition, the specific types of decisions made via examination of these results will be discussed. Extension of the user-relevant verification concept to other types of verification studies – to achieve the goal of providing user-specific and actionable information – will also be considered.

Title (joint keynote): MET and MesoVICT - Tools and Data for the Application and Testing of Established and New Spatial Verification Methods

Authors: Barbara Brown[1] and Manfred Dorninger[2]
1 = National Center for Atmospheric Research, Boulder, CO, USA
2 = University of Vienna, Vienna, Austria

The Model Evaluation Tools (MET) community verification package provides a wide range of tools for evaluation of forecasts. Development of MET was initiated at a time of enthusiasm and interest in new verification approaches and tools, born partly out of the need for methods that could provide meaningful information about forecasts from new high-resolution mesoscale models and partly as a result of new developments in verification that could meet those needs. This presentation considers the historical context associated with MET's initiation as well as the evolution of MET since 2007, with a particular focus on the spatial verification methods that have been a key focus of MET development and application. MET's expansion to a more complete package (incorporating new infrastructures and tools), and its global future, will also be discussed.

The focus of MesoVICT (MesoScale Verification Intercomparison over Complex Terrain) is the application, assessment, and enhancement of the capability of spatial verification methods over complex terrain. Deterministic and ensemble forecasts at near-convection-resolving and convection-permitting resolution are evaluated. Test cases include additional variables beyond precipitation, such as wind and temperature. The cases represent interesting meteorological events in a region of complex terrain over the Alps. Objective gridded analysis fields and observations from a very dense stations network are provided. The aim is to refresh the guidance for the end user on the best use of spatial verification approaches. On the other hand, this analysis will potentially identify shortcomings in existing methods. The talk summarizes the outcomes that shall lead to benefits for verification users and verification method developers.

Title: Decomposition of Verification Scores for Deterministic Continuous Forecasts: Insights from Image Quality Assessment

Authors: Dominique Brunet[1] and Gabrielle Gascon
[1] = Environment and Climate Change Canada

Many verification scores for deterministic continuous forecasts can be expressed in term of means, variances and correlation between forecasts and observations. Examples of verification scores with such decomposition include the Mean Squared Error (MSE), the Nash-Sutcliffe Efficiency (NSE), and the Kling-Gupta Efficiency (KGE) and its variant. These verification scores bear some resemblance to the Structural Similarity (SSIM) index, an image quality assessment measure popular for digital pictures. The SSIM index and its variant, the Multi-Scale SSIM (MS-SSIM), combine means, variances and correlation computed on a local neighborhoods. Scores in the SSIM family are compared with other scores decomposition, including the Fractional Skill Score (FSS), which is seen to only depend on the mean values (i.e. fractions) in a local neighborhood.

We are interested in unifying these different scores into a common framework and obtain insights from previous work on the study of the Structural Similarity index image quality assessment measure. We discuss on (i) spatial decomposition of scores and their aggregation into a global score, (ii) terms for additive or multiplicative errors, (iii) robustness to noise and other small perturbations, (iv) parameter tuning and multi-objective scores, (v) distance-metric properties, and (vi) score optimization for post-processing tasks. We conclude by looking beyond scores based on means, variances and correlation to address two issues: (i) spatial errors, (ii) non-Gaussian error distribution. We use data from a validation experiment of the Integrated Multi-SatellitE Retrievals for the Global Satellite Mission (GPM-IMERG) against a Probabilistic Quantitative Precipitation Estimation (PQPE) over Canada to illustrate the concepts.

Title: The role of model calibration and verification techniques for sub-seasonal weather regime forecast skill

Authors: Dominik Büeler [1], Julian F. Quinting, Jan Wandel, Christian M. Grams
1= Karlsruhe Institute of Technology, Germany

The continuous increase of computational power and improvement of numerical weather prediction systems in recent decades has allowed extending the operational weather forecast horizon into sub-seasonal time scales (10 – 60 days). On these scales, quasi-stationary, persistent, and recurrent large-scale flow patterns, so-called weather regimes, explain most of the regional surface weather variability in the midlatitudes and are thus of primary interest in sub-seasonal forecasting for the respective region.

Here, we investigate how different model calibration as well as verification techniques affect the (still) only moderate to weak skill in predicting Atlantic-European weather regimes on sub-seasonal time scales. In a first part, we elucidate how forecast skill is affected by classical compared to flow-dependent model calibration techniques. In a second part, we demonstrate that the forecast skill horizon is larger for predicting the occurrence of a weather regime for lead-time-dependent verification windows rather than at individual lead time steps. Applying such techniques to operational sub-seasonal forecasting systems can contribute to the envisaged extension of the forecast skill horizon into weeks 3 and 4.

Title: Do short-range convection-permitting ensembles lead to more skilful probabilistic rainfall forecasts over Tropical East Africa ?

Authors: Carlo Cafaro[1], Beth Woodhams, Thorwald Stein, Cathryn E. Birch, Stuart Webster, Caroline Bain, Andrew Hartley, Samantha Clarke, Samantha Ferret, Peter Hill
1 = University of Reading, UK

This study examines a novel set of limited-area convection-permitting ensembles (CP-ENS) run by the UK Met Office covering the tropical East Africa domain, for 24 cases in the period April-May 2019. CP-ENS and its parent global ensemble with parametrized convection (Glob-ENS) are evaluated against rainfall estimates derived from satellite observations (GPM-IMERG). CP-ENS represents better the diurnal cycle than Glob-Ens, with heavy rainfall amounts generally overestimated by CP-ENS. Probabilistic forecasts of heavy rainfall were then generated and verified using a neighbourhood approach using both ensembles and their deterministic counterparts. Fractions skill score (FSS) and areas under the ROC curve metrics (AUC) were implemented to assess the spatial skill and discriminating ability respectively. FSS results show that CP-ENS has useful skill for scales greater than 100 km. Skill decreases with lead time and shows diurnal signals, especially for CP forecasts. Pairwise comparisons between the different forecasts reveal that CP-ENS is generally the most skilful for both 3-h and 24-h accumulations, followed by the CP deterministic counterpart. AUC results are consistent with these findings. Also, CP-ENS was found to be under-spread both spatially and for domain average rainfall, in agreement with other studies. Overall, the results of this study shows the benefits and limitations of CP-ENS over tropical Africa and should stimulate further research and development, including best practice guidance to local forecasters.

Title: Representativeness issues in verification practices

Authors: Barbara Casati[1], Francois Lemay[2], Nelson Shum[2], Thomas Haiden[3]
1 = Meteorological Research Division, Environment and Climate Change Canada
2 = Canadian Meteorological Service, Environment and Climate Change Canada
3 = European Centre for Medium-Range Weather Forecasts, UK

Day-to-day changes in surface variables are partially driven by the general circulation, but they are also strongly influenced by (model sub-tile) local characteristics. Surface variable verification is therefore strongly affected by representativeness issues, more so than upper-air verification. In this work we illustrate two examples in which representativeness dominates the error signal.

The first example addresses the coarse representation of topography in the model, and shows that cold temperature biases in mountainous terrain are often due to the mis-match between model height versus station elevation. To address the mis-match, model temperatures are then adjusted to station elevation by applying a standard atmosphere lapse-rate, as well as an inversion lapse-rate in calm and clear nights. The lapse-rate adjustment, despite quite simplistic, significantly affects the verification results and ranking between high and low resolution models.

The second example pertains to local effects driven by the presence of water bodies and different land-types, and illustrates that some specific model sub-tile components (such as the land-portion of the tile versus the water-portion of the tile) often better represent the weather observed at land-based versus water-based stations.

Interpretation of verification results should always be performed in full awareness of representativeness issues: novel verification approaches need to be developed in order to identify and separate the representativeness error from the model error.

Title: **Incorporating the perspective of user evaluation into the creation of a new early warning system.**

Authors: MPP. Julia Chasco
Head of Meteorology and Society Department – National Meteorological Service of Argentina.

The creation of user-oriented services is common practice in different sectors, both public and private. The innovation of services based on scientific knowledge has been transitioning for years to new forms and methodologies that try to incorporate the view of the beneficiary of these solutions through the collection of needs from a critical and interdisciplinary perspective.

The HIW/WWRP-endorsed Alert.Ar Project has inaugurated in the National Meteorological Service of Argentina a new way of addressing the needs of users who use weather and climate services to mitigate the impacts of severe events. Since 2015, a small group of social sciences junior researchers have been working to identify opportunities to improve the NMS's early warning system, and in five years they have become a stable body within the NMS. The knowledge acquired in that experience not only managed to collect the users' voice in detail but also to establish new interdisciplinary ways of working that today are reflected in a recently pre-launched early warning system that incorporates these voices.

What do our users say about the information services that science provides? What preconceptions does the scientific system hold when it comes to providing solutions to citizens? What are the interdisciplinary methodologies that can be incorporated into meteorological services to provide useful services? How do we incorporate the knowledge of the user who makes decisions based on our service, considering that these decisions can save lives? Or better, where would we start if we could radically change an early warning system?

This presentation will attempt to collect the most relevant aspects of the work carried out by the Meteorology and Society department from Project Alert.Ar in 2015 to the establishment of a new early warning system in 2020 and provide the knowledge acquired on strengthening science-based services created from transdisciplinarity.

Title: A verification method for sub-daily rainfall intensity in short-range forecast

Authors: Haoming Chen[1], Jian Li, Nina Li
1 = Chinese Academy of Meteorological Sciences, China

With the continuous advancement of urbanization, the society has higher requirements for fine-scale precipitation forecast. The concern is not only the daily rainfall intensity, but also when and where it will rain during the day. To improve the fine-scale quantitative precipitation forecast (QPF) in short-range, more objective evaluation metrics for sub-daily rainfall variation need to be take into consideration. Most of the operational verification for QPF are based on accumulated rainfall amount, while the sub-daily variation has not been extensively verified.

A method to linearly evaluate hourly rainfall frequency-intensity distribution is applied to verify the QPF from operational regional convective-permitting models (CPM) in China Meteorological Administration (CMA). Two parameters are used to quantitatively illustrate the model capability in reproducing the percentage of weak and heavy precipitation. By comparing the spatial distribution and diurnal variation of the parameters, it is shown that the overestimation of the weak rainfall over the steep terrains such as the eastern slope of Tibetan is still a common bias in operational CPM forecast. The false prediction of afternoon weak rainfall dominates the bias of diurnal variation. These results indicate that the method well demonstrates the capability of sub-daily QPF, and it is now applied in the operational regional model verification system in CMA.

Title: A symmetric spatial verification method for sea ice edges

Authors: Angela Cheng[1], Barbara Casati[3], Jean-Francois Lemieux[3], Bruno Tremblay[2], and Adrienne Tivy[1]

1 = McGill University and Canadian Ice Service, Environment and Climate Change Canada
2 = McGill University, Canada
3 = Meteorological Research Division, Environment and Climate Change Canada

The sea ice edge demarcates where the sea ice pack ends and open water begins. It is an important feature for northern travel and for sea ice models. A common method for spatial verification of forecasted sea ice edges that attempts to maintain spatial structure is the Hausdorff distance. The Hausdorff distance is not a symmetric spatial verification method—comparing a forecast against an observation can yield a different result than comparing an observation against a forecast. We present a similar, but symmetric, spatial verification method as an alternative to the Hausdorff distance. By comparing this method against the Hausdorff distance, we can discuss the merits of symmetry: what additional information can symmetry provide, and conversely what information can be gained from lack of symmetry?

Title: Measuring uncertainty in sea ice concentration observations in Canadian Ice Service ice charts

Authors: Angela Cheng[1], Barbara Casati[3], Adrienne Tivy[1], Tom Zagon[1], Jean-Francois Lemieux[3] and Bruno Tremblay[2]
1 = McGill University and Canadian Ice Service, Environment and Climate Change Canada
2 = McGill University, Canada
3 = Meteorological Research Division, Environment and Climate Change Canada

Ice charts generated by national ice services are generated by trained analysts who analyze satellite imagery to estimate sea ice concentrations. These ice concentration estimates are used to report current conditions to mariners, to initialize sea ice models, to develop sea ice climatologies, and are used for validation of automated algorithms for calculating sea ice concentration in remotely sensed imagery. Of interest is a) analyst's accuracy in estimating the ice concentration, and b) variability between analyst estimates. Few studies on this topic have been done due to operational time constraints, resulting in little standard in measuring and reporting of this uncertainty between national ice services. In this study we used Krippendorff's alpha, which has been used in communication studies, and employ it for measuring variability between analysts. We showed that analyst responses had high agreement with one another but that collectively, analysts overestimate ice concentrations. The uncertainty may have downstream implications for numerical modelling and sea ice climatology.

Title: The Mediterranean, Black and Marmara Seas analysis and forecasting physical systems: validation methodology and quality assessment

Authors: Emanuela Clementi[1], Stefania Angela Ciliberti, Mehmet Ilicak, Ivan Federico, Ivano Barletta, Elisaveta Peneva, Alí Aydogdu, Jenny Pistoia, Eric Jansen, Romain Escudier, Leonardo Lima, Massimiliano Drudi, Alessandro Grandi, Laura Stefanizzi, Rita Lecci, Sergio Cretí, Francesco Palermo, Vladyslav Lyubartsev, Anna Chiara Goglio, Diana Azevedo, Murat Gunduz, Salvatore Causio, Simona Masina, Nadia Pinardi, Giovanni Coppini

1 = Centro Mediterraneo sui Cambiamenti Climatici, Italy

The Mediterranean (MedFS) and Black Sea (BSFS) operational forecasting systems, developed in the context of the Copernicus Marine Environment and Monitoring Service (CMEMS), produce analyses and 10-days forecasts of the main physical variables: temperature, salinity, sea level, currents, mixed layer depth (the MedFS at 4.5 km resolution and 141 vertical levels, the BSFS at 3 km and 31 vertical levels). The hydrodynamic core is based on NEMO (Nucleus for European Modelling of the Ocean) model, coupled to OceanVar data assimilation method to assimilate in-situ and satellite observations. A near real time operational skill assessment is provided (http://medfs.cmcc.it/ and http://bsfs.cmcc.it/) to monitor the product quality, showing analysis and forecast field maps at different depths and weekly validations of model analysis compared with available observations. To allow for an optimal interface between the MedFS and the BSFS, a new high resolution model for the Marmara Sea including the Bosporus and Dardanelles Straits has been developed using the System of HYdrodynamics Finite Element Modules (SHYFEM). It uses high resolution unstructured mesh up to 50 metre resolution in the horizontal to resolve the Turkish Straits and 93 geopotential coordinate levels in the vertical. A pre-operational simulation has been carried out and validated using the seasonal in situ observational data. The operational version of the Marmara Sea model will provide lateral open boundary conditions to the MedFS and BSFS systems. The focus of this work is to present the modelling systems and their validation assessment including comparison with in-situ and satellite observational datasets.

Title: Evaluating the representation of precipitation variability patterns over South America in sub-seasonal predictions

Authors: Caio Coelho[1], Felipe Andrade, Marisol Osman, Mariano Alvarez, Carolina Vera and Iracema Cavalcanti
1 = CPTEC/INPE, Brazil

Anticipated sub-seasonal precipitation information (expected conditions for next 4 weeks) is important for activities planning in various application sectors. South America sub-seasonal precipitation is strongly influenced by key modes of variability, with well defined spatial patterns, representing the main sources of predictability in this time scale. Therefore it is important to investigate how well sub-seasonal prediction models represent these precipitation patterns over South America to have an assessment of their reproducibility as well as to better understand the mechanisms behind the actual prediction quality. To achieve this goal Empirical Orthogonal Function (EOF) analysis is used for identifying these spatial patterns, together with the corresponding principal component time series, in observations and a selection of Sub-Seasonal to Seasonal (S2S) Prediction Project models, including the multi-model ensemble mean of the investigated models. The employed methodology also allows quantifying the contributing of the identified patterns to the actual S2S models precipitation retrospective prediction quality, assessed through the association attribute, as well as to the estimated potential predictability.

Title: Methods and Tools Used to Verify Convection-Allowing Model Guidance at the NCEP/Environmental Modeling Center

Authors: Logan C. Dawson[1], Geoffrey S. Manikin, Perry C. Shafran, Binbin Zhou, Christopher MacIntosh, Jason J. Levit

1= NOAA/NCEP/EMC, USA

The Verification, Post-Processing, and Product Generation Branch at the National Centers for Environmental Prediction's Environmental Modeling Center (NCEP/EMC) has worked in recent years to consolidate and bolster its verification capabilities to meet the testing and evaluation requirements for proposed upgrades to NCEP's production suite. These efforts include transitioning all verification packages to utilize the Model Evaluations Tools (METplus) and incorporating methods and metrics following community standards outlined for NOAA's emerging Unified Forecast System (UFS). Highlights of verification development focused on NCEP's high-resolution, convection-allowing models (CAMs) will be presented; a particular focus will be placed on methods used for verifying guidance for severe convective environments and storms. Presently, a METplus-based verification package has been implemented into the workflow for the high-resolution UFS Short Range Weather application. This effort ensures all developers will use consistent methods and metrics to verify upper air, surface, precipitation, aviation, and severe weather guidance. Details regarding the probabilistic and neighborhood-based methods used to verify simulated reflectivity and updraft helicity guidance will be discussed. Additionally, a novel, quasi-Lagrangian approach to verifying severe weather guidance in NCEP/Storm Prediction Center convective outlook areas, forecast regions where severe convective storms and hazards are anticipated, is being used to demonstrate the utility of such regime-based verification. Verification results for severe weather guidance offer a proof of concept supportive of extending this regime-based verification approach to other phenomena, such as fire weather and excessive rainfall, with similarly defined outlook areas.

Title: Measuring the impact of a new snow model using surface energy budget process relationships

Authors: Jonathan J. Day[1], Gabriele Arduini[1], Irina Sandu[1], Linus Magnusson[1], Anton Beljaars[1], Gianpaolo Balsamo[1], Mark Rodwell[1] and David Richardson[1]
1= European Centre for Medium-Range Weather Forecast, UK

Energy exchange at the snow-atmosphere interface in winter is important for the evolution of temperature at the surface and within the snow, preconditioning the snowpack for melt during spring. This study illustrates a set of diagnostic tools that are useful for evaluating the energy exchange at the Earth's surface in an Earth System Model, from a process-based perspective, using in-situ observations. In particular, a new way to measure model improvement using the response of the surface temperature and other surface energy budget (SEB) terms to radiative forcing is presented. These process-oriented diagnostics also provide a measure of the coupling strength between the incoming radiation and the various terms in the SEB, which can be used to ensure that improvements in predictions of user relevant properties, such as 2m temperature, are happening for the right reasons. Correctly capturing such process relationships is a necessary step towards achieving more skilful weather forecasts and climate projections.

These diagnostic techniques are applied to assess the impact of a new multi-layer snow scheme in the European Centre for Medium-Range Weather Forecasts'-Integrated Forecast System at two high-Arctic sites (Summit, Greenland and Sodankylä, Finland). A previous study showed that it will enhance 2m temperature forecast skill across the northern hemisphere in boreal winter compared to forecasts with the single layer model, reducing a warm bias. In this study we use the diagnostics to show that the bias is improved for the right reasons.

Title (keynote): Forecast quality assessment for operational climate prediction

Authors: Francisco Doblas-Reyes[1], Carlos Delgado, Nube González-Reviriego, Carlo Lacagnina, Andrea Manrique, Albert Soret, Verónica Torralba, Deborah Verfaillie
1 = Barcelona Supercomputing Center, Spain

The knowledge transfer between research and operations is key for climate forecast operations to fulfil the increasing number of identified user requirements. Climate forecast operations can benefit from many of the topics dealt with by forecast quality assessment research: the appropriate verification of a forecast product against observational datasets including incorporating uncertainty measures in forecast quality attributes, the importance of dealing with observational uncertainty, the definition of meaningful benchmarks. Forecast quality assessment also guides the approaches to construct more reliable and skilful products from, for instance, multi-models, or lagged ensembles. This talk will illustrate how these aspects play a role in climate forecast operations, including examples from a range of time scales from sub-seasonal to decadal prediction. It will also highlight the need for standards to document forecast systems and, in particular, their performance in a coherent way with other data types relevant in the generation of climate information. The talk will address the challenges that climate services have to face to integrate forecast quality information of operational products in formats understandable by a wide range of users.

Title: Measure of Forecast Challenge and Predictability Horizon Diagram Index for Ensemble Models

Authors: Jun Du[1], Binbin Zhou and Jason Levit
1= Environmental Modeling Center, NCEP/NWS/NOAA

This study proposed two new verification metrics to quantify the forecast challenges that a user faces in decision making when using ensemble models. The measure of forecast challenge (MFC) combines forecast error and uncertainty information together into one single score. It consists of four elements: ensemble mean error, spread, nonlinearity, and outliers. The cross correlation among the four elements indicates that each element contains independent information. The relative contribution of each element to the MFC is analyzed by calculating the correlation between each element and MFC. The biggest contributor is the ensemble mean error, followed by the ensemble spread, nonlinearity, and outliers. By applying MFC to the predictability horizon diagram of a forecast ensemble, a predictability horizon diagram index (PHDX) is defined to quantify how the ensemble evolves at a specific location as an event approaches. The value of PHDX varies between 1.0 and -1.0. A positive PHDX indicates that the forecast challenge decreases as an event nears (type I), providing creditable forecast information to users. A negative PHDX value indicates that the forecast challenge increases as an event nears (type II), providing misleading information to users. A near-zero PHDX value indicates that the forecast challenge remains large as an event nears, providing largely uncertain information to users. Unlike current verification metrics that verify at a particular point in time, PHDX verifies a forecasting process through many forecasting cycles. Forecasting-process-oriented verification could be a new direction in model verification. The sample ensemble forecasts used in this study are produced from the NCEP global and regional ensembles.

Title (keynote): Overview of the HIWeather User-Oriented Evaluation Task Team Activities

Authors: Beth Ebert[1] (Task Team Lead) and the HIWeather User-Oriented Evaluation Task Team
1 = Bureau of Meteorology, Melbourne, Australia

The WWRP High Impact Weather (HIWeather) project aims to "promote cooperative international research to achieve a dramatic increase in resilience to high impact weather, worldwide, through improving forecasts for timescales of minutes to two weeks and enhancing their communication and utility in social, economic and environmental applications." The improvement in the quality and value of forecasts can occur in many different parts of the value chain, which extends from the observation and modelling of high impact weather and associated hazards to understanding and predicting the risk of societal and environmental impacts, creating and communicating effective warnings, and community benefit resulting from taking appropriate action in response to the warning.

The HIWeather User-Oriented Evaluation Task Team seeks to measure the quality and value of the warning value chain through projects that explore different elements of the value chain and the information that connects the various users along that chain. These include exploring how to use spatial verification methods to measure the accuracy of high resolution ensemble forecasts, using non-traditional observations to assess high impact weather forecast performance, understanding how to use ensemble forecasts to effectively warn for the risk of flood associated with heavy rainfall in tropical cyclones, surveying weather services on the impact-based warning practices, comparing intended and actual response to severe weather warnings, and estimating avoided losses from effective warnings. The other three webinar speakers in the session will describe user-oriented evaluation of tropical cyclone and fire forecasts, and incorporating the perspective of user evaluation into Argentina's new early warning service.

A new HIWeather flagship project will use value chain approaches to evaluate the end-to end warning chain for case studies of high impact weather collected in a database.

Title: Nonparametric Permutation Procedures for Evaluating Climate Model Accuracy

Authors: Joshua P. French[1] and Piotr S. Kokoszka
1 = University of Colorado, USA

There are many stakeholders interested in the behavior of future climate. E.g., insurance companies need to assess how future climate may impact the cost and risk of potential claims. Future climate behavior is typically projected using large-scale computer-based simulations, which are generated for both historical and future time periods under different scenarios. It is natural to assess the trustworthiness of these projections by determining whether the simulated historical climate matches what was observed. Comparisons of observed and modeled climate behavior often focus on central tendencies, which overlooks other important distributional characteristics related to extreme quantiles and variability. We present two nonparametric permutation procedures, standard and stratified, for assessing the accuracy of climate models. Both permutation procedures make only weak assumptions about the underlying data structure, encouraging their application in variety of contexts. By making only slightly stronger assumptions, the stratified procedure dramatically strengthens the ability to detect a difference between the observed and climate model behavior. The proposed procedures allow researchers to identify potential model deficiencies over space and time for a variety of distributional characteristics, providing a more comprehensive assessment of climate model accuracy, which will hopefully lead to further model refinements. Application will be made to state-of-the-art data from the North American Coordinated Regional Climate Downscaling Experiment (NA-CORDEX).

Title: MEC – A web-based tool for multi-model weather forecasting evaluation comparison

Authors: Garcia[1], J.R.M, Figueroa S.N., Rozante, J.R.
1 = CPTEC/INPE, Brazil

A highly interactive web-based system has been developed for processing and viewing statistical forecast verification of multi-NWPSs over South America and it is called MEC (Model Evaluation Comparator). In addition to the intrinsic aspect of the high disseminating potential that web-based applications have, MEC allows visualising the forecast verification from different ways.

Results are shown according to the choices of a set of pre-established scopes in its web-interface, such as models, meteorological variable, observational dataset, spatial and temporal domain, forecasting lead-time, and statistical metric. To do that, a preprocessing phase in which a meteorological variable of all involved regional and/or global forecasting models plus its correspondent observation are scaled to a common spatial domain and stored in NetCDF files. Next, the daily-basis forecast verification is done by combining all the pre-established scopes. All of these combinations are processed in order to generate known continuous and categorical metrics. Data for the Taylor diagram, Performance diagram, frequency data for histograms and JPEG image files of the whole South America domain with respect to forecasting, observation and bias are also generated.

The interface explores the power of the R language and its graphical system to automatically fit charts on the screen and also averaging results according to the period. Taylor and Performance diagrams, scores cards, and frequency histograms are also present in this under development verification system.

Title: Spatial Verification: A New Spatial Alignment Error Summary

Author: Eric Gilleland[1]
1 = NCAR, USA

Many spatial forecast verification methods are largely centered on the issue of timing or displacement errors, and therefore focused on informing about spatial alignment errors between the forecast and observation fields, including: (i) amount of overlap, (ii) distance between misses and false alarms, (iii) differences in shapes of features (aka objects, event areas) either individually (features-based techniques) or on the whole (e.g., distance-based measures). The methods vary in complexity but often one or a few summary measures are required in order to facilitate large numbers of cases. A recent investigation of several distance-based summaries revealed specific challenges for these measures and, in some cases, fundamental flaws. A new spatial alignment summary measure is proposed, here, that is shown to handle all of the shortcomings of these other summaries. It requires the user to choose one parameter that scales the results according to the type of information sought; in most cases, however, half the square of the domain size provides informative results. Time-permitting, the talk will touch on incorporation of intensity error handling in the spatial context and introduce two modifications of the proposed summary measure that also incorporate the intensity component.

Title: Appraisal of Challenging WeAther foREcasts (AWARE) in COSMO

Authors: Gofa[1] F., Bundel A, Tesini M.S., Marsigli C., Mazur A., Linkowska J., Hoff M., Boucouvala D., Duniec G., Cattani D., Tatarinovich E., Muraviev A.
1 = Hellenic National Meteorological Service, Greece

The importance of accurate forecasting of challenging weather occurrences is obvious. The increased demand to provide accurate forecasts of extreme weather especially during highly convective weather events leads to the question to how objectively to evaluate forecasts. Within this study, a number of forecast methods and evaluation approaches that are linked to high impact weather are tried providing the COSMO Community with an overview and recommendations as to how such situations should be handled.

The phenomena that are studied are mainly intense precipitation and thunderstorms with the associated lightning activity and visibility range restrictions. The ability of various commonly used verification measures to represent forecast quality is assessed as no single score exists that addresses all properties of such relatively rare events. In addition to this, spatial approaches are proved particularly useful for the evaluation of very high resolution forecasts. Measures and approaches such as FSS, DIST, CRA, SAL, MODE (MMI) are used with special attention to the correct matching between the forecasted element and the observed quantity that is representing of the phenomenon of interest. Finally, special attention is given to the necessary post-processing methods and their ability to represent high impact weather, compared to direct model output. This is accomplished, in particular, by exploring approaches associated to machine learning such as multi-linear regression (MLR), adaptive least squares (A-RLS) and/or artificial neural networking (ANN) techniques with an effort to propose appropriate ways of representing and communicating HIW forecast for decision making.

Title: Verification of Quantile Forecasts – A Journey

Authors: Deryn Griffiths[1], Robert Taggart, Michael Foley, Nicholas Loveday, Alistair McKelvie, Ben Price
1 = Australian Bureau of Meteorology

The Bureau of Meteorology issues various forecasts for daily rainfall including the mean, the median, other quantiles, and the chance of exceeding various thresholds.

We verify the chance of exceeding a given threshold using the Brier Score. However, our initial attempts to assess forecasts such as the 90th percentile of daily rainfall was very simply showing the proportion of times the observation exceeded the forecast. This showed some measure of the reliability of the forecasts but gave no overall measure of its skill. A climatological forecast would score perfectly on this measure.

Recently, we learnt about consistent scoring functions for single value quantile forecasts. For forecast x predicting the α quantile of the forecast distribution, and observation y, we can score the forecast as follows:

$$\begin{array}{ll} \alpha|x - y| & \text{if } x \leq y, \\ (1-\alpha)\,|x - y| & \text{if } x > y. \end{array}$$

For the median forecast, the score is essentially the mean absolute error.

By introducing this score, we will be able to track improvement in forecasts of a particular quantile and to compare two forecasts of the same event in a meaningful way. To compare the whole forecast distribution, we use the (Continuous) Ranked Probability Score. However, verifying a point of the forecast probability distribution is important if that value is a prominent aspect of one's forecast service, or known to be used by a client for a particular decision.

This talk will showcase techniques for verifying a rainfall probability distribution, including point values from the distribution, and discuss the decisions being informed by the verification.

Title (keynote): Process-oriented verification

Author: Thomas Haiden[1]
1 = European Centre for Medium-Range Weather Forecast, UK

The purpose of verification in numerical weather prediction (NWP) ranges from creating high-level information for management decisions (administrative), through assessing the value of forecasts for users (user-oriented), to model development (often process-oriented). The main goal of process-oriented verification is a better understanding of model issues. Here, a ‘process’ can be a specific atmospheric phenomenon, e.g. the formation of nighttime cold air pools in basins, or a specific model component, e.g. the parameterization of surface-atmosphere coupling. A key method of tackling process-related questions is the use of a large array of concurrent, and often co-located, complementary observations such as at supersites or in field experiments. Another important aspect is verification methodology. Stratification of the data, awareness about the statistical properties of the metrics used, and creative visualization of verification results (for pattern recognition by humans or AI) can be effective tools. This talk presents examples from ongoing work in NWP and related areas to illustrate what works well in process-oriented verification and diagnostics, as well as some of the issues that have been encountered.

Title: Measuring Performance, Skill and Accuracy in Operational Oceanography : Overview of approaches proposed by the GODAE/Ocean Predict Intercomparison and Validation Task Team

Authors: Fabrice Hernandez[1], Gregory Smith[2], and the Ocean Predict Intercomparison and Validation Task Team (IV-TT)
1 = IRD/LEGOS (France) & UFPE/DOCEAN (Brasil)
2 = Meteorological Research Division, Environment and Climate Change Canada

Operational oceanography is now established in many countries, focusing on global, regional, and/or coastal areas, and targeting different aspects of the « blue », « white » or « green » ocean. There are nowadays a large variety of interests and users, with different disciplines and level of expertise. Validation and verification of operational products aims at quantifying the level of confidence of this variety of ocean products. Since 1998 the Validation and Intercomparison Task Team (IV-TT) as part of GODAE and now the Ocean Predict international framework, has been acting in order to federate scientific development of validation and verification approaches, facing many challenges: Ocean models reaching the submesoscales not adequately observed ; many products available for a given ocean variable, not always consistent ; real time forecasting systems challenged by reanalyses ; more complex operational systems based on coupled models (ocean, ice, atmosphere, biogeochemichal) … In parallel, the global ocean observing system is continuously updated with additional satellites, with innovative sensors, with improved integration of the global, regional and coastal in-situ observing capabilities. This presentation aims to provide an overview of the challenges and approaches addressed by the IV-TT to offer a common international framework for evaluating operational oceanography performances, like the Class1-4 framework.

Title: "Twin-analysis" verification: a new verification approach that alleviates pitfalls of "own-analysis" verification when applied to short-range forecasts

Authors: Daisuke Hotta[1], Takashi Kadowaki[2], Hitoshi Yonehara[2], Toshiyuki Ishibashi[1]
1 = Meteorological Research Institute, Japan Meteorological Agency
2 = Japan Meteorological Agency

In operational NWP, forecast verification against analysis from the same experiment is part of the standard evaluation practice. This "own-analysis" verification is beneficial in providing complete spatial coverage but is known to suffer from the data dependency issue when applied to short-rage forecasts where the inevitable positive correlation between the forecast and analysis errors leads to overly optimistic verification scores. This issue is particularly problematic when a new development involves assimilation of new observations since the more observations we assimilate, the less correlated the background and the analysis tend to be, leading to apparent degradation in the verification score which makes interpretation of verification results delicate and difficult. To alleviate this problem, we propose to perform "twin-analysis" verification in which we produce "twin analyses" by running an independent cycle using the same NWP system as the one used to produce the forecasts, but initializing from an independent first guess at the beginning of the cycling period and then we verify the forecasts against these twin analyses. This way the error correlation between the forecasts and analyses should be reduced while preserving the statistical properties of the analyses, hopefully enabling a clearer interpretation of verification. In this talk we will report the results of comparison between "twin-analysis" and "own-analysis" verification scores obtained for the JMA's global NWP system. The two scores disagree up to two days, suggesting that "own-analysis" verification can be unreliable for such short ranges.

Title: Fostering International Collaboration Through a Unified Verification, Validation, and Diagnostics Framework - METplus

Authors: Tara Jensen[1], Marion Mittermaier, Jason Levit, Elizabeth Satterfield, Evan Kuchera, Louisa Nance
1 = National Center for Atmospheric Research, USA

Verification and validation activities are critical for the success of modeling and prediction efforts at organizations around the world. Having reproducible results via a consistent framework is equally important for model developers and users alike. The Model Evaluation Tools (MET) was developed over a decade ago and expanded to the METplus framework with a view towards providing a consistent platform delivering reproducible results. The METplus system is an umbrella verification, validation and diagnostic tool for use by thousands of users from both US and international organizations. These tools are designed to be highly flexible to allow for quick adaption to meet additional evaluation and diagnostic needs. A suite of python wrappers have been implemented to facilitate a quick set-up and implementation of the system, and to enhance the pre-existing plotting capabilities. Recently, several organizations within the National Oceanic and Atmospheric Adminstration (NOAA), the United States Department of Defense (DOD), and international partnerships such as Unified Model (UM) Partnership led by the Met Office have adopted the tools for their use both operationally and for research purposes. Many of these organizations are also now contributing to METplus development, leading to a more robust and dynamic framework for the entire earth system modeling community to use. This presentation will provide an overview of METplus and how it is being used in across multiple scales and applications. It will highlight examples of METplus applied to verification and validation efforts throughout the international community to address a range of temporal (hourly forecasts to subseasonal-to-seasonal) and spatial scales (convection allowing to mesoscale, regional to global, tropical to cryosphere to space).

Title: Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited

Authors: Timo Dimitriadis[2,3], Tilmann Gneiting[1,2], Alexander Jordan[2]
1 = Karlsruhe Institute of Technology, Germany
2 = Heidelberg Institute for Theoretical Studies, Germany
3 = University of Hohenheim

A probability forecast or probabilistic classifier is reliable or calibrated if the predicted probabilities are matched by ex post observed frequencies, as examined visually in reliability diagrams. The classical binning and counting approach to plotting reliability diagrams has been hampered by a lack of stability under unavoidable, ad hoc implementation decisions. Here we introduce the CORP approach, which generates provably statistically Consistent, Optimally binned, and Reproducible reliability diagrams in an automated way. CORP is based on non-parametric isotonic regression and implemented via the Pool-adjacent-violators (PAV) algorithm - essentially, the CORP reliability diagram shows the graph of the PAV-(re)calibrated forecast probabilities. The CORP approach allows for uncertainty quantification via either resampling techniques or asymptotic theory, furnishes a new numerical measure of miscalibration, and provides a CORP based Brier score decomposition that generalizes to any proper scoring rule. We anticipate that judicious uses of the PAV algorithm yield improved tools for diagnostics and inference for a very wide range of statistical and machine learning methods.

Title: Taking the wind-induced undercatch of solid precipitation into account in the verification process.

Authors: Morten Køltzow[1], Barbara Casati[2], Thomas Haiden[3], Teresa Valkonen[1]
1 = MET Norway, 2 = Environment and Climate Change Canada, 3 = ECMWF

All precipitation observations have associated uncertainties, making it difficult to quantify the true forecast quality. One of the largest uncertainties is due to the wind-induced undercatch of solid precipitation gauge measurements. We show how this observation error impacts the verification of precipitation forecasts for Norway where solid precipitation is dominant during winter.

First, the forecasts are compared with high-quality reference measurements (less undercatch) and with more simple measurement equipment, commonly available and used in verification (substantial undercatch), at the WMO SPICE Haukeliseter observation site in Norway. Then the verification is extended to include all Norwegian observation sites; 1) stratified by wind speed, since calm (windy) conditions experience less (more) undercatch; 2) by applying transfer functions, which convert measured precipitation to what would have been measured with high-quality equipment with less undercatch, before the forecast-observation comparison is performed.

Results show that the wind-induced undercatch of solid precipitation has a substantial impact on verification results. Applying transfer functions to adjust for wind-induced undercatch of solid precipitation gives a more realistic picture of true forecast capabilities. In particular, estimates of systematic forecast biases are improved, as well as metrics like correlation, RMSE, ETS and SEEPS, but to a lesser degree. However, uncertainties associated with applying transfer functions are substantial and need to be taken into account in the verification process, which is discussed. In addition, the application of transfer function requires information on hourly precipitation, wind speed and temperature not always available from observations. Some strategies on how to tackle this are discussed.

Title: Regional Class 4 verification of the Canadian operational ice-ocean prediction systems

Authors: Yvonnick Le Clainche[1], Greg Smith[2], Jinshan Xu[1], Fraser Davidson[1], Yimin Liu[2], Frederic Dupont[2]
1 = Department of Fisheries and Oceans, Canada
2 = Environment and Climate Change Canada

Ice-ocean analysis and forecasting systems have been developed and operationally implemented under the Canadian Operational Network of Coupled Environmental Prediction Systems (CONCEPTS), an inter-departmental initiative involving Environment and Climate Change Canada (ECCC), Fisheries and Oceans Canada (DFO) and the Department of National Defense (DND). A DFO Service Desk for Operational Oceanography (SeDOO) was established to be the DFO hub for the application of operational ice-ocean prediction systems. As part of its mission to support real-time monitoring and analysis, SeDOO has notably undertaken the management and the update of the verification and evaluation tools based on Class-4 metrics defined by GODAE OceanView (GOV), now Ocean Predict, for Sea Surface Temperature (SST), Sea Level Anomaly (SLA), temperature and salinity profiles and sea ice concentration. The analysis and forecasts of the Global Ice-Ocean Prediction System (GIOPS) are near real-time evaluated against observations and compared with other models that participate in the GOV international benchmarking. Class-4 metrics are also calculated for the new Regional Ice-Ocean Prediction Systems (RIOPS), whose the geographical domain covers the North Atlantic, the Arctic and the North-East Pacific. Statistics of the difference between the observations and the “model equivalents” are calculated into various sub-areas allowing to better assess the quality of the GIOPS and RIOPS forecasts and their skills in key regions of interest for Canada, such as the North-East Pacific, the North-West Atlantic or the Canadian Arctic. This is necessary to guide and enhance operational CONCEPTS system uses in those areas.

Title: Seasonal Forecast Skill of ENSO Teleconnection Maps

Authors: Nathan Lenssen[1], Lisa Goddard, and Simon Mason
1 = International Research Institute for Climate and Society (IRI), Columbia University, USA

The El Niño-Southern Oscillation (ENSO) is the dominant source of seasonal climate predictability. This study quantifies the historical impact of ENSO on seasonal precipitation through an update of the global ENSO teleconnection maps of Mason and Goddard (2001). Many additional teleconnections are detected due to better handling of missing values and 20 years of additional, higher quality data. These global teleconnection maps are used as deterministic and probabilistic empirical seasonal forecasts in a verification study. The probabilistic empirical forecast model outperforms climatology in the tropics demonstrating the value of a forecast derived from the expected precipitation anomalies given the ENSO phase. Incorporating uncertainty due to SST prediction shows that teleconnection maps are skillful in predicting tropical precipitation up to a lead time of four months. The historical IRI seasonal forecasts generally outperform the empirical forecasts made with the teleconnection maps, demonstrating the additional value of state-of-the-art dynamical-based seasonal forecast systems. Additionally, the probabilistic empirical seasonal forecasts are proposed as reference forecasts for future skill assessments of real-time seasonal forecast systems.

Title: Evaluating probabilistic forecasts with scoringRules

Authors: Alexander Jordan[2], Fabian Krüger[1], Sebastian Lerch[1]
1 = Karlsruhe Institute of Technology, Germany
2 = Heidelberg Institute for Theoretical Studies, Germany

Probabilistic forecasts in the form of probability distributions over future events have become popular in several fields including meteorology, hydrology, economics, and demography. In typical applications, many alternative models and data sources can be used to produce probabilistic forecasts. Hence, evaluating and selecting among competing methods is an important task. The scoringRules package for R provides functionality for comparative evaluation of probabilistic models based on proper scoring rules, covering a wide range of situations in applied work. I will discuss implementation and usage details, and presents case studies from meteorology.

Title: Understanding the link between ensemble mean error variance, spread-error ratio, mean error and the CRPS

Authors: Martin Leutbecher[1] and Thomas Haiden[1]
1 = European Centre for Medium-Range Weather Forecast, UK

A statistical model is introduced with the aim of better understanding the characteristics of the continuous ranked probability score (CRPS). The model assumes a homogeneous Gaussian (hoG) distribution for the joint distribution of forecast and observations. This hoG model permits to express the expected CRPS analytically as a function of the variance of the error of the ensemble mean, the mean error of the ensemble mean and the ensemble variance. Moreover, the hoG model comes with an elegant decomposition of the CRPS into reliability and resolution components. This expression for the CRPS can be applied to verification statistics of any ensemble and provides an approximation of the actual CRPS.

The usefulness of the approximation to predict the sample mean CRPS and to predict sample mean changes of the CRPS between different forecast-observation distributions is investigated with operational medium-range ensemble forecasts. The hoG approximation could be exploited for new diagnostics in verification software used by NWP developers routinely. The diagnostic may help NWP developers (or users) to better explain differences between the CRPS of two ensemble forecasts in terms of the respective differences in mean error, ensemble mean error variance and ensemble variance. The diagnostics requires little additional computational resources compared to the alternative of verifying postprocessed versions of the ensemble forecasts. Therefore, the diagnostics could be applied easily to a large set of variables that is scrutinized as part of the model development process.

Title: Measure of Forecast Challenge and Predictability Horizon Diagram Index for Ensemble Models

Authors: Jason Levit[1] and Geoffrey Manikin
1 = Environmental Modeling Center, NCEP/NWS/NOAA

In recent years, the Environmental Modeling Center (EMC) re-organized, and established the new Verification, Post-Processing, and Product Generation Branch (VPPPGB). The new branch was created as part of a decision to recognize the increasing community emphasis on improving the verification and evaluation of environmental prediction systems. As the global weather enterprise continues to evolve towards more nuanced, detailed, and increasingly complex data high resolution prediction systems, the need for verifying and evaluating has increased substantially, especially to support the National Weather Service's Impact-Based Decision Support Services (IDSS) structure and the emerging Unified Forecast System. The new VPPPG Branch is therefore designed to help EMC organize towards supporting new verification and evaluation efforts. This presentation will describe the structure of the new branch, and discuss how state-of-the-art verification and evaluation methodologies are used to examine the performance of EMC's prediction systems. For example, EMC is moving towards exclusively using the Model Evaluation Tools (METplus) software system for all models, and through the EMC reorganization, the software will be used to establish the branch as the independent verification component of EMC modeling efforts. A discussion of these topics, future projects, and desired community partnerships with the new branch will be included in the presentation.

Title: Exploring Spatial Distributions of Systematic Errors in the NCEP's Global Ensemble Precipitation Forecast Products

Authors: Yan Luo[1] and Jason Levit
1 = Environmental Modeling Center, NCEP/NWS/NOAA

Despite recent progresses in numerical weather prediction, the ensemble precipitation forecasts are still prone to systematic biases, remaining a challenge for NWP model guidance products. Understanding such a persistent problem, how much spatial variations of systematic errors exist in global ensemble model precipitation forecast product forecast has been an ongoing and interesting research topic. Assessing such performance of precipitation forecast is important for future research-to-operations activities and for forecasters to better understand NWP output. Moreover, bias correcting precipitation forecast is hopefully a necessary post-processing step in the operational global ensemble forecasting.

In this study, 24-hour precipitation forecasts in NCEP's GEFS operational and bias-corrected products are evaluated for a selected period and verified at different lead-times and thresholds from a spatial variability of view. Various comparisons are also employed to demonstrate the usefulness and effectiveness of this bias-correction approach. Findings are evaluated by conventional metrics such as the mean value, mean error, frequency, and frequency bias and will be expanded more in the future study.

We will present preliminary results to identify and characterize of the systematic errors in the forecast products. The results can be used to provide recommendations for model developers to improve the products. Then we will discuss plans to add more diagnostic studies such as seasonal variability of biases, focusing on efforts to provide useful information for identifying model limitations and weaknesses, and provide diagnostic metrics to improve model and ensemble forecast performance.

Title (keynote): Understanding medium-range forecast errors from a synoptic-dynamic perspective

Authors: Linus Magnusson[1]
1 = European Centre for Medium-Range Weather Forecast, UK

Medium-range weather forecasts have undergone significant improvement since the advent of global forecasts 40 years ago. The improvements originate from enhanced observation systems, improved data assimilation together with improved models. However, occasionally, forecasts experience very low scores, and such episodes are often referred to as ’forecast busts’ or ’dropouts’.

Understanding the root causes of very poor forecasts and the processes involved in the error growth is essential but difficult. Different techniques for tracking sources of errors include manual error tracking, ensemble sensitivity and nudging (relaxation) experiments toward the analysis. While the error tracking and ensemble sensitivity can be applied on standard model output, nudging experiments need access to run the model. Another technique is to use the same initial conditions but different models to understand the relative impact of errors in initial conditions and model formulation. Here we use 2 sets of forecasts with the same model but different initial conditions and 2x2 sets with same initial conditions. These pairs of forecasts give the opportunity to disentangle the impact of errors from initial conditions and from the model, and also to get a better understanding of some systematic errors.

In the presentation we will discuss results using different diagnostic tools both for busts in mid-latitude and Arctic forecasts, and how the cases of large errors connect to different weather features. We will discuss the advantages and disadvantages with the different tools. The results will both give guidance for forecast system developments and highlight situations where forecasters need to be more cautious.

Title: The Model Evaluation Group at the Environmental Modeling Center

Authors: Geoffrey S. Manikin[1], Alicia M. Bentley, Logan C. Dawson, Shannon R. Shields, Christopher MacIntosh, Philippe Papin, and Jason J. Levit
1 = Environmental Modeling Center, NCEP/NWS/NOAA

The Model Evaluation Group (MEG) was formed at the Environmental Modeling Center (EMC) in 2012 to completely revamp the EMC approach of evaluating the performance of operational and parallel models. Prior to the formation of the MEG, validation of upgrades to systems was based on a small number of key metrics, and there was no organized vetting process. There was also limited interaction and communication between model customers and developers.

The role of the MEG has expanded in recent years to lead the formal assessment of proposed upgrades to modeling systems. This involves an examination of real-time and (when available) retrospective model runs, using an approach of both objective and subjective measures. In recent years, the Environmental Modeling Center (EMC) re-organized, and established the new Verification, Post-Processing, and Product Generation Branch (VPPPGB). The MEG works closely with the Verification staff of the VPPGB to identify and document key verification metrics and then with the user community to examine how verification metrics translate into day-to-day forecast maps. In addition to working with the user community to vet proposed model implementations, the MEG helps document systematic model biases to be targets for improvement going forward. This is also accomplished by reviewing and comparing model performance for high-impact events.

This presentation will describe the evolution of the MEG within EMC and its expanding role. It will discuss new approaches for identifying and defining key metrics for different applications of the Unified Forecast System (UFS) and how the MEG will continue to identify needed areas of improvement within UFS components through the evaluation of existing models, new verification metrics, and interactions with customers and stakeholders.

Title: Choices in the verification of S2S forecasts

Authors: Andrea Manrique-Suñén[1], Nube Gonzalez-Reviriego; Verónica Torralba; Nicola Cortesi; Francisco J. Doblas-Reyes
1 = Barcelona Supercomputing Center, Spain

Sub-seasonal predictions bridge the gap between medium-range weather forecasts and seasonal climate predictions. This time scale is crucial for operations and planning in many sectors such as energy and agriculture. In order for users to trust these predictions and efficiently make use of them in decision making, the quality of predicted near surface parameters needs to be systematically assessed. However, the methodology to follow in a probabilistic evaluation of sub-seasonal predictions is not trivial.

This study aims to offer an illustration of the impact that the verification setup might have on the calculation of the skill scores, thus providing some guidelines for sub-seasonal forecast evaluation. For this, several forecast verification setups to calculate the fair ranked probability skill score for tercile categories have been designed. These setups use different number of samples to compute the fair RPSS as well as different ways to define the climatology, characterised by different time periods to average (week or month).

These setups have been tested evaluating 2 m temperature in ECMWF-Ext-ENS 20-years hindcasts against the ERA-Interim reanalysis. Results show that in order to obtain a robust skill score several start dates need to be employed. It is also shown that a constant monthly climatology over each calendar month may introduce spurious skill score associated with the seasonal cycle. A weekly climatology bears similar results to a monthly running window climatology, however the latter provides a better reference climatology when bias adjustment is applied.

Title (keynote): Observations for high-impact weather and their use in verification

Author : Chiara Marsigli[1]
[1] = Deutscher Wetterdienst, Germany

The verification of high-impact weather requires a different approach to the traditional objective verification process normally used for meteorological variables involved in the occurrence of high-impact weather phenomena (e.g., precipitation, temperature, wind). For the purposes of verifying forecasts of high-impact weather, “traditional” observations often do not permit characterization of the phenomenon of interest, and therefore do not provide a good reference for objective verification. This talk presents a review of new observations, or more generically quantities which can be considered as reference data or proxies, which can be used for the verification of high-impact weather phenomena. The options are many and varied, from remote sensing datasets, datasets derived from telecommunication systems including cell phones, data collected from citizens, reports of impacts and claim/damage reports from insurance companies. Only two phenomena are addressed: thunderstorms and fog, selected as representative of potentially high-impact weather of interest to users. In particular, it is described what is needed to transform these different sources of information about high-impact weather phenomena into objective data to be used for performing a statistical verification. Example taken from the scientific literature are provided.

Title: Helping the agricultural food chain make better choices: The value of seasonal climate forecasts for European maize production.

Authors: Alberto Ceccacci, Stefano Materia[1]
[1] = Centro Mediterraneo sui Cambiamenti Climatici, Italy

In Southern Europe, more frequent drought phenomena will severely affect highly water-demand crops such as maize, whose yield is deeply influenced by water deficits during flowering stage. Allowing farmers, buyers and insurers to know in advance the risk of a dry summer, long-term forecasts can provide valuable information to a variety of economic agents in the maize market. In this study, we decide to compute the expense reduction that these actors can get by trusting a seasonal forecast of drought delivered in March, which enables them to adopt precautionary measures before summer. The decision scenario is described using the cost-loss model, both in its traditional version and a newly developed rendering, built to fit the three users of the multi-model seasonal forecasts delivered by Copernicus C3S. After defining a drought severity index for the spatial extent and intensity of droughts, we applied a categorical score to convert the probabilistic forecast into a deterministic forecast, in order to determine the correctness of the prediction. It emerges that- depending on their cost-loss ratio, buyers, farmers and insurers benefit from a maximum reduction in their average expense of approximately 30% of what would be obtained through a decision process based on standard approaches. These results suggest that economic analysis can boost the uptake of climate information in Southern Europe, thus promoting the development of more advanced techniques for the elaboration of long-term forecasts and contributing to the adaptation capacity of one of the regions that is mostly exposed to global warming.

Title: Outcome-conditioned Decompositions of Proper Scores

Authors: Keith Mitchell[1] and Chris A.T. Ferro
1 = Exeter University, UK

A popular way to diagnose the performance of forecasts is to decompose proper scores in to measures of reliability, resolution and uncertainty. This is based on factorising the joint distribution of forecasts and outcomes by conditioning on the forecasts. Such decompositions, however, can fail to distinguish forecasters with different powers of discrimination and so separate measures of discrimination are required. Murphy and Winkler (1987) showed that a measure of discrimination can be obtained from the Brier score by means of a complementary decomposition in which the joint distribution is conditioned on the outcomes. We show that such outcome-conditioned decompositions are available for all proper scores.

Title: Using MODE and MODE TD to investigate the evolution of the 2019 Chlorophyll-a bloom season in the North West European Shelf region

Authors: Marion Mittermaier[1], Rachel North, Jan Maksymczuk and Christine Pequignet
1 = UK Met Office

The feature-based verification methods MODE and MODE Time Domain (TD), commonly used for atmospheric model applications, were applied to Chlorophyll-a (Chl-a) concentration forecasts from the AMM7 North West European Shelf Seas model, and compared against gridded satellite observations of Chl-a concentration from the Copernicus Marine Environmental Monitoring Service (CMEMS) catalogue. Two forms of quantile mapping were used to deal with a diagnosed concentration bias before the objects were matched to ensure that the analysis of spatial properties was not dominated or obscured by the presence of a concentration bias. Despite this, forecast objects were found to be too large though forecast objects tended to be found in the right sort of locations, though not necessarily at the right time. Analysing the space-time objects showed that the bloom is modelled too late in the forecast model by around a month. This work was part of the Copernicus Marine Environment Monitoring Service (CMEMS) High-resolution Verification and Evaluation (HiVE) project, and represents, to our knowledge, the first application of spatial verification methods to ocean forecasts.

Title: Verification of Tanzanian Meteorological Authority Severe Weather impact-based forecasts

Authors: Hellen E. Msemo[1], Cathryn E. Birch, Tamora James, Beth J. Woodhams, Andrea L. Taylor, Andrew J. Dougill, Mark Richardson
1 = University of Leeds, UK

Recent advances in numerical weather prediction (NWP) have increased forecast precision, reliability and lead time, with at least some skill in forecasting extreme weather events in the tropics. Whilst weather warnings are regularly issued in Tanzania, the country continues to suffer from the impacts of weather-related disasters. Between 2000 and 2019, severe weather accounted for approximately 69% of disasters in Tanzania. There is a clear need to better understand the value of weather warning and advisory information and how this information can be better used in the decision-making processes of those receiving advisories and warnings. This work seeks to better understand the entire chain of the forecasting process, from the generation of forecasts to the dissemination of weather products to the end user. In this work we evaluate Tanzanian Meteorological Authority weather warnings and advisories for heavy rains at lead times of 1-5 days against satellite precipitation retrievals. The forecasts are issued in map and text format within a pdf document that is issued daily. A number of technical steps were required to extract the relevant information from the pdf documents. We answer the following questions: i) Which thresholds for heavy rainfall are most relevant? ii) How does the accuracy of the warnings vary with location? iii) How many warnings were misses or false alarms? iv) How accurate was the severity of the warning?

Title: Sub-Seasonal Forecast Skill: When, Where and How To Find It?

Authors: Á.G. Muñoz[1], C.A.S. Coelho, S.J. Mason, A.W. Robertson, K. Pegion, and F. Vitart

1 = International Research Institute for Climate and Society (IRI), Columbia University, USA

Recent research has highlighted the potential for improving predictive skill at the sub-seasonal timescale, which could be the basis for enhanced, actionable forecasts for climate services involving water and disaster management, health, energy and food security. Projects such as WMO's World Weather and World Climate Research Programme's Subseasonal-to-Seasonal Prediction Project (S2S) and NOAA's SubX have made available extensive databases with both hindcasts and almost-realtime forecast at this timescale. Lead times are long enough that much of the information in the atmospheric initial conditions is lost, but at the same time are too short for other sources of predictability (e.g., ocean boundary conditions) to have a strong in-fluence in skill. Presently, sub-seasonal skill is still limited beyond 2-3 weeks, and in general uncalibrated forecasts cannot be used to develop climate services. An obvious alternative is to make use of a variety of robust statistical calibration methods –also known as Model Output Statistics, MOS– available for other timescales, such as the seasonal one. Nonetheless, different methods have different advantages and disadvantages, depending on which forecast attribute to focus on. Here, as a benchmark, we first analyze the spatio-temporal variability of predictive skill in uncalibrated models, and then discuss how local (gridbox-by-gridbox) and non-local (pattern-based) calibration models enhance or decrease skill in different regions of the world.

Title: Detecting over-confidence in weather forecasts

Authors: Kenric Nelson[1] and Harold Brooks[2]
1 = Photrek, USA, 2 = NOAA National Severe Storms Laboratory, USA

The accuracy of weather forecasts has often been measured by the mean-square average of the probability forecasts, beginning with Brier. While Shannon entropy and its associated logarithmic score, are known to provide a stronger theoretical grounding for forecast assessment, the steep penalties for probabilities near zero are considered overly stringent. The methodologies of proper scores which subtract the biases in metrics such as the mean-square average are considered adequate for evaluation of weather forecasts. Unfortunately, the result is systemic over-confidence in the reports relied on for forecasting.

We show how a local score based on a generalization of information theory can detect over-confidence. The method utilizes the generalized mean of the probabilities assigned to actual events as a measure of performance. The power of the generalized mean defines the degree of risk tolerance bias. The full spectrum of metrics forms a Risk Profile in which negative powers are highly sensitive to outliers and thus useful in detecting forecasts which are over-confident.

A comparison between an algorithm's performance and the estimated source of uncertainty provides a quantitative visualization of the degree of over-confidence. The Risk Profile's utility in analyzing weather forecasts is demonstrated with two sets of precipitation forecasts in the Oklahoma, US area. The results show that while the one- and two-day forecasts give an accurate assessment of the uncertainty, the six- and seven-day forecasts are highly over-confident. A simple anchoring of the long-range forecasts to the long-term base rate would dramatically improve the accuracy of the forecasts.

Title: A new skill score for quantifying the uncertainty in multi-category precipitation forecasts

Authors: Yawei Ning[1], Xiaogang Shi, Guohua Liang, Bin He, Wei Ding, Huicheng Zhou
1 = Dalian University of Technology, China

The uncertainty analysis is critical for multi-category precipitation forecasts because uncertainty is an inherent ingredient in the hydrometeorological forecasting. Traditional verification methods, however, are mainly focused on the metrics for the accuracy rather than the uncertainty in multi-category precipitation forecasts. Therefore, this study proposed a new skill score for quantifying the uncertainty in multi-category precipitation forecasts based on the entropy theory. The skill score is defined as a ratio of reduced uncertainty in actual precipitation after receiving a certain category precipitation forecast information to the initial uncertainty of actual precipitation. By using the new entropy theory based skill score, the four precipitation forecast products from China Meteorological Administration (CMA), European Centre for Medium-Range Weather Forecasts (ECMWF), National Centers for Environmental Prediction (NCEP) and United Kingdom Meteorological Office (UKMO) were verified in the Hunhe River Basin, Northeastern China.

Title: Reference forecast of sea-ice edge using damped persistence of probability anomaly

Authors: Bimochan Niraula[1] and Helge Goessling
1 = Alfred Wegner Institute, Germany

Recent advancement in dynamical sea-ice models have enabled weather agencies to forecast sea-ice conditions at sub-seasonal to seasonal timescales. Using the S2S dataset, the ice-edge output of various forecasting centers was compared against reference forecasts to assess the predictive skill of the models. However, the simplest types of reference forecasts – persistence of the initial state and climatology – do not exploit the observations optimally and thus lead to overestimation of forecast skill. For spatial objects such as the ice-edge location, the development of damped-persistence forecasts that combine persistence and climatology in a meaningful way poses a challenge. With this motivation, we have developed a probabilistic reference forecast method that combines the climatologically derived probability of ice presence with initial anomalies of the ice edge. We have tested and optimized the method based on minimization of the Spatial Probability Score, and compared it to the output from other models in the S2S dataset. The resulting reference forecasts provide a challenging benchmark to assess the added value of dynamical forecast systems.

Title: Using diagnostics from calculating verification scores to identify systematic errors

Authors: Rachel North[1], Marion Mittermaier, Sean Milton

1 = Met Office, United Kingdom

Satellite observations of precipitation enable assessment of weather forecast models in locations where traditional observation sources don't, notably over the ocean regions. The Stable Equitable Error in Probability Space (SEEPS) score, modified for use with a satellite-derived daily precipitation climatology dataset (using the Tropical Rainfall Measuring Mission 3B42 research product), was used to evaluate a new global configuration of the Met Office Unified Model (UM), which has been presented previously. Following on from this work, to try and understand signals coming from the overall analysis, extra spatial diagnostic capabilities have been produced with the intermediate stages of the SEEPS calculation, in order to try and identify sources of systematic error. We present these diagnostics derived from the UM data set in question and show the resulting conclusions, which include a vast improvement in the handling of light daily precipitation over the Tropics in the new configuration.

Title: Evaluating the Impact of Planetary Boundary Layer, Land Surface Model, and Microphysics Parameterization Schemes on Cold Cloud Objects in Simulated GOES-16 Brightness Temperatures

Authors: Sarah M. Griffin[1], Jason A. Otkin, Sharon E. Nebuda, Tara L. Jensen, Patrick S. Skinner, Eric Gilleland,Timothy A. Supinie, Ming Xue
1 = University of Wisconsin-Madison, USA

Infrared brightness temperatures (BTs) from the GOES-16 Advanced Baseline Imager are used to examine the ability of several microphysics and planetary boundary layer (PBL) schemes, as well as land surface models (LSM) and surface layers, to simulate upper-level clouds. Cloud objects are identified using the Method for Object-Based Diagnostic Evaluation (MODE) and analyzed using object-based and pixel-based metrics. Object-based metrics include the Object-Based Threat Score and Mean-Error Distance. Pixel-based metrics include the mean absolute error and mean bias error (MBE) for matched objects, after removing displacement between objects. Objects are identified using either a fixed BT threshold of 235 K or the 6.5th percentile of for each model configurations. Analysis of the MODE-identified cloud objects shows that the microphysics scheme had the largest impact on the accuracy of the simulated cloud field. The Thompson scheme produced the most accurate cloud objects, whereas the Morrison-Gettelman scheme had the lowest accuracy due to too many cloud objects covering too large of an area compared to the observations. Changing the PBL scheme from Mellor-Yamanda-Nakanishi-Niino (MYNN) to Shin-Hong or Eddy-Diffusivity Mass-Flux resulted in slightly lower accuracy, however these changes resulted in configurations which better reproduced the number of observation cloud objects and slightly reduced the high MBE that occurred when using the MYNN PBL. Changes to the LSM are found to have an accuracy ranging between the changes to the microphysics scheme and the PBL, however changing the LSM and surface layer resulted in too many objects as the forecast progressed.

Title: User-oriented verification of the DWD climate prediction website

Authors: Andreas Paxian[1], Katja Reinhardt, Klaus Pankatz, Katharina Isensee, Kristina Fröhlich, Barbara Früh
1= German Weather Service (DWD), Germany

DWD provides operational seasonal and decadal predictions since 2016 and 2020, respectively. We plan to present these predictions together with post-processed ECMWF monthly forecast products on the DWD climate prediction website www.dwd.de/climatepredictions. In March 2020 it has been published with decadal predictions; other time scales will follow. It offers maps, time series and tables of ensemble mean and probabilistic predictions of temperature and precipitation for different regions. Further user-oriented or extreme variables and statistical downscaling will follow. The user-oriented evaluation and design has been developed in close cooperation with users from various sectors at workshops of the German MiKlip project and will be consistent across decadal, seasonal and monthly time scales. Climate predictions are displayed in combination with their skill. The MSESS and fairRPSS are used to evaluate the skill of climate predictions in comparison to reference predictions, e.g. 'observed climatology' or 'uninitialized climate projections' (which are applied until now as an alternative to climate predictions). Significance is tested via bootstraps. Within the 'basic climate predictions' section, a user-oriented traffic light indicates whether regional-mean climate predictions are significantly better (green), not significantly different (yellow) or significantly worse (red) than reference predictions. Within the 'expert climate predictions' section, prediction maps show per grid box the prediction itself (via the color of dots) and its skill (via the size of dots representing the skill categories of the traffic light). The co-development of this climate prediction website improves the way it is understood and can be applied by users in their daily work.

Title: On the benefits of AMDAR Observation Profiles for Forecast Validation

Authors: Ralph Alvin Petersen[1] and Timothy J. Wagner
1 = University of Wisconsin-Madison, Cooperative Institute for Meteorological Satellite Studies (CIMSS), Space Science and Engineering Center (SSEC)

In addition to being far more cost effective than other data sources, wind/temperature observations from commercial aircraft are critical to operational Numerical Weather Prediction (NWP) on global (3rd highest impact of all data types) and regional scales. The value of the high-quality moisture reports from the Water Vapor Sensing System (WVSS) may be less well recognized but can be much larger locally, as moisture changes often occur at much smaller temporally, vertically, and spatially scales. Comparisons between nearly 140 aircraft and precisely co-located raob observations across CONUS show agreement of 0.5 g/kg, with minimal biases, and suggest that WVSS measurement are as or more accurate than raob moisture measurements. Regional NWP studies show that WVSS profiles have equal or greater influence than raobs in 1-2 day forecasts.

Tropospheric AMDAR ascent/descent profiles have been used to validate other asynoptic observations, including EUMETSAT MetOp IASI moisture retrievals and can be performed throughout the entire day, unlike raob comparisons at only 00 and 12 UTC. Results clearly demonstrate the strengths and weakness of the satellite observation in the free atmosphere and boundary layer.

AMDAR profiles are also being used to assess the impact of various new observations in fine-scale regional data assimilation and forecasts. Hourly or sub-hourly time series of AMDAR profiles at multiple geographical locations provide a long-desired opportunity to contrast in-situ observations with predicted atmospheric variations and tendencies at temporal and vertical resolutions unavailable from other operational observing system. This paper describes ongoing NWP validations and enhancements planned for the future.

Title: Using Integrated Ice Edge Error (IIEE) and Spatial Probability Score (SPS) to assess spread-error relationships in an ensemble sea ice forecast

Authors: Andrew (Drew) Peterson[1] and Greg Smith
1 = Environment and Climate Change Canada

The deterministic Integrated Ice Edge Error (IIEE) and the Spatial Probability Score (SPS), an area integrated Brier Score have been put forward as useful metrics to score deterministic and ensemble sea ice forecasts respectively. While these metrics have been used to provide comparative scores of sea ice forecasts between centres and as measures of system advancement, here we use them in the context of the newly coupled Environment and Climate Change Canada (ECCC) Global Ensemble Prediction System (GEPS), where comparison scores are not an option for our first set of medium range ensemble sea ice forecasts. Instead, we show how to make use of differences between the IIEE and SPS to measure the spread error characteristics of our system, providing examples of where the probabilistic information of our ensemble system might prove useful.

Title: Deep Learning for the Verification of Synoptic-scale Processes in NWP and Climate Models

Authors: Julian F. Quinting[1], Christian M. Grams[1], Jan Wandel[1]
1= Karlsruhe Institute of Technology, Germany

Physical processes on the synoptic scale are important modulators of the large-scale extratropical circulation. In particular, rapidly ascending air streams in extratropical cyclones, so-called warm conveyor belts (WCBs), have a major impact on the Rossby wave pattern and are sources and magnifiers of forecast uncertainty. Thus, an adequate representation of WCBs is desirable in NWP and climate models. Most often, WCBs are defined as Lagrangian trajectories that ascend in two days from the lower to the upper troposphere. This Lagrangian approach has advanced our understanding of the involved processes significantly. However, the calculation of trajectories is computationally expensive and requires data at high spatio-temporal resolution so that systematic evaluations of the representation of WCBs in NWP and climate models are missing. In this study, we present a novel framework that aims to predict the inflow, ascent, and outflow phases of WCBs from instantaneous gridded fields. A UNet-type Convolutional Neural Network (CNN) is trained using a combination of meteorological parameters as predictors. Validation against a Lagrangian-based dataset confirms that the CNN model reliably replicates the climatological frequency of WCBs as well as their footprints at instantaneous time steps. With its comparably low computational costs we propose that the new diagnostic may be applied to systematically verify WCBs in large datasets such as ensemble reforecast or climate model projections. Our diagnostic demonstrates how deep learning methods may be used to advance our fundamental understanding of synoptic-scale processes that are involved in forecast uncertainty and systematic biases in NWP and climate models.

Title: User decisions, and how verification based the utility of these decisions could guide developments in probabilistic forecasting

Authors: Mark Rodwell[1], John Hammond, Sara Thornton, David Richardson
1 = European Centre for Medium-Range Weather Forecasts, UK

We investigate how users combine objective probabilities with their own subjective feelings when deciding how to act on weather forecast information. Results are based on two scenarios investigated at a Live Science event held by the Royal Meteorological Society. When deciding whether to go to the beach with the possibility of warm, dry weather, we find that users attempt to identify their 'Bayes Action': the one which minimises their expected negative feeling or utility. Key factors are the 'thrill' of a nice day at the beach and the 'pain' of coping with, for example, children in wet weather, and the costs of travel. The users' threshold probabilities for deciding to go to the beach thus approximately define their distribution of cost/loss ratios. This is used to calculate a 'User Brier Score' (UBS): a measure of the overall utility to society, and which could be used to guide forecast system development. When applied to operational ensemble forecasts issued by the European Centre for Medium-Range Weather Forecasts (ECMWF) over the period 1995–2018, the UBS tends to be higher (i.e., worse) than the Brier Score, largely because users tended not to exhibit high cost/loss ratios. When deciding whether to leave a campsite in the face of potentially dangerous gales, users try to find a balance between the 'regret' of serious injury and the 'pain' of spoiling an enjoyable holiday. Some users decide to stay even at high probabilities of serious consequences – partly due to a lack of experience. On the other hand, forecasts suffer from 'complete misses' – where probabilities of zero are accompanied by non-negligible outcome frequencies. These dominate the overall Brier Score. The frequency of complete misses halved over the period 1995–2018: a welcome improvement for users who do wish to avoid danger at low probabilities.

Title: A New METplus-based Verification System for the Global Forecast System (GFS)

Authors: Mallory Row[1] and Jason Levit
1 = NOAA/NCEP/EMC, USA

The Verification, Post-Processing and Product Generation (VPPPG) branch at the Environmental Modeling Center (EMC) has worked in recent years to unify its verification systems following a decision by the Next Generation Global Predictions System (NGGPS) program. This decision originally involved the use of the Model Evaluation Tools (MET) and was later expanded to include the METplus authoritative umbrella repository, which is comprised of MET, METviewer and METexpress (two database and display systems), and a set of python wrappers around all these tools. METplus is maintained and developed by the Developmental Testbed Center (DTC) in Boulder, Colorado. A new METplus-base verification package has been developed at EMC to verify the Global Forecast System (GFS). The new package is a collection of shell and python scripts that set up a user's environment, gets data, and calls METplus to create verification statistics and graphics. The package was originally developed to replicate the capabilities of the previously used, in-house developed verification package, and, as a part of future work, it will expand upon these original capabilities. The package has been incorporated into the GFS workflow, and verification statistics produced from this new package have been used in the evaluation of the newest version of the GFS proposed for implementation in early 2021. This package is available on GitHub allowing it to be checked out and run by users across the broader United States numerical weather modeling community.

Title: Nowcasting verification in Argentina's Weather Service (SMN)

Authors: Facundo San Martino[1], Sebastián Pérez, Pablo Irurzun, Pedro Lohigorry, Ramón de Elía[1]
1 = Servicio Meteorológico Nacional, Argentina

The Weather Service of Argentina is responsible for issuing short-term weather warnings mostly related to severe convective events. The Nowcast division monitors the situation and identifies areas in which severe weather is already present or is likely to appear in the near future, with the aim of warning the public, media, government and emergency managers. Monitoring is carried out with the help of weather radars, satellite imagery, lightning detectors and a (sparse) official network of weather stations. Forecasters need to be as certain as possible that inferences based on remote sensing correspond well with the situation on the ground. Ground observations have, therefore, a very important role, both in triggering forecasts and also in the verification of information from other sources. In order to complement a sparse official network of weather stations, the Nowcasting division has striven to utilize information available from other agencies, corporations, passionate weather amateurs, and more general communication of participants in social networks. In this presentation we will discuss how we gather this information and use it to confirm events such as first touchdown in areas threatened with severe weather, and also how this information, along with that compiled from media reports, is later sifted and used to verify the evolution of the severe weather as predicted by forecasters. Focus will be put on the challenges of dealing with sources of varying credibility and having large spatial density variation, and the difficulty of arriving at a fair forecasting skill score.

Title: Forecasting spatial structure of local precipitation extremes

Author: Bent Sass[1]
[1] = Danish Meteorological Institute (DMI), Denmark

In view of an increasing interest to predict extremes, e.g. in relation to daily weather forecasts, a spatial verification scheme has been developed which compares the match of local extremes of a forecast field and a corresponding analysis field. The concept is illustrated for accumulated precipitation analyzed spatially in DMI and compared directly to forecasted precipitation accumulation in the same grid. The scheme computes four separate scores between zero and 1 for identified local extremes (maxima and minima) of analysis and forecast fields respectively. The four separate comparisons are done in local neighborhoods around respectively: analyzed maximum point(s) (compared with forecasted maximum in neighborhood), forecasted maximum (compared with analyzed maximum in corresponding neighborhood). A corresponding computation in neighborhoods is done for observed and forecasted minima respectively. The degree of match is defined by a score function giving score=1 for perfect match between forecast an analysis and zero for large deviation between the two. In this way the scheme defines 4 scores. The scheme has been tested successfully in idealized tests, a forecast case of convection and in simulations of operational conditions. Differences and similarities with popular spatial verification schemes such as FSS and SAL have been studied.

Title: Verification of Air Quality Predictions Using METplus

Authors: Perry C. Shafran[1], Ho-Chun Huang, Jianping Huang, Edward Strobach, Partha Bhattacharjee, and Jeffery McQueen
1 = NOAA/NWS National Air Quality Forecasting Capability

The NOAA/NWS National Air Quality Forecasting Capability (NAQFC) has been used to provide numerical guidance of ground-level ozone and particulate matter with diameters less than 2.5 micrometers (PM25). In NAQFC, the United States (U.S.) Environmental Protection Agency (EPA) Community Multiscale Air Quality Modeling (CMAQ) system is driven by meteorological models, compared to parallels using FV3 model output. The Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT) is driven by the NCEP (National Centers for Environmental Prediction) operational models. Integration of global aerosol prediction at NOAA/NWS is based on the Goddard Chemistry Aerosol Radiation and Transport (GOCART) scheme into the Unified Forecast System (UFS) begun by including it into one member of the Global Ensemble Forecast System (GEFS-Aerosol).To examine the predictive performance of these variables, it is important to evaluate the NAQFC performance at different regions and time periods. Verification uses Model Evaluation Tools (MET), driven by METplus python wrappers. Traditional surface ozone and PM25 observations from the EPA (Environmental Protection Agency) AirNow network, and satellite observed variables such as the aerosol-optical depth have been added for the air quality verification. Also, GEFS-Aerosols has been extensively evaluated against a number of ground observations, analysis and satellite observations. Other features developed in METplus - including the creation of masked data to examine, for example, biases related to land surface and their impacts on air quality - are also showcased. This presentation compares VSDB verification of ozone and PM25 with MET verification and describes the performance matrix for NAQFC model evaluation.

Title: New operational measure to assess extreme events using site-specific climatology

Authors: Michael A Sharpe[1], Clare Bysouth and Philip Gill
1= UK Met Office

Work to implement a measure to assess how well extreme weather events are forecast is nearing completion at the UK Met Office. This methodology assesses post-processed, site-specific data for extremes of maximum daily temperature, minimum daily temperature and hourly wind speed. The Threshold Weighted Continuous Ranked Probability Skill Score (Gneiting and Rajan, 2011) and the Threshold Weighted Mean Absolute Error are used for this analysis, enabling a direct comparison between probabilistic and deterministic forecast performance. In each case, a monthly site-specific CDF is derived for the threshold weighting function using a bootstrapping procedure based on a chosen percentile from the a 30-year climatological PDF. To help with communication to the general public, some carefully chosen percentiles (in addition to the standard ones used to evaluate extremes) are used for this assessment. These evaluate the ability to forecast the most extreme event that should be expected to occur at a given site and month during an n-year period. The effect of climate change on extremes is incorporated into the measure by recalculating the monthly, site-specific 30-year climatological PDF (and consequently the corresponding threshold weighted CDF for each chosen percentile) every year.

Title: A complementary measure to assess temporal uncertainty within Terminal Aerodrome Forecasts

Authors: Michael A Sharpe[1] and Andre Lanyon
1= UK Met Office

Terminal Aerodrome Forecasts (TAFs) are a widely accepted international form of aviation forecast used for airport and flight planning procedures at all major airports. A new verification methodology (Sharpe et al, 2016) has recently been developed and made operational at the UK Met Office to assess the skill associated with TAFs at UK airports. This methodology is based on the definitions devised and published by the World Meteorological Organisation and International Civil Aviation Organisation. These definitions allow forecasters to use probabilistic, deterministic and temporal uncertainty terms. However, these terms can be used excessively to hedge performance and even though the resulting forecast conforms to the rules stipulated by the WMO/ICAO a less uncertain forecast would usually be more useful to the user. Therefore, this presentation outlines a complementary uncertainty-penalising TAF verification methodology developed using probability theory. Its use alongside the established Sharpe et al measure will help to identify whether identified good performance is a result of actual forecaster skill or hedging via the excessive use of terms containing forecast uncertainty.

Title: Analysis of Regional Sector Low-Skill Events in Recent Operational GFS Forecasts

Authors: Shannon R. Shields[1], Travis J. Elless, and Daryl T. Kleist
1 = IMSG at NOAA/NWS/NCEP/EMC, USA

While forecast skill continues to improve with model upgrades, there are still occasional periods where the forecast skill is significantly reduced, especially on regional scales. Regional sector low-skill events were categorized based on 120-h 500-hPa height anomaly correlation coefficient (ACC) and root mean square error (RMSE) and diagnosed to determine causes of operational Global Forecast System (GFS) forecast error since June 2019. For this study, five-day ACC and RMSE were calculated using the most recent 0000 UTC initialization date in the GFS archive for each regional sector. The regional sectors included: Eastern North America/United States, Western North America/United States, North America in general, Eastern Pacific, Western Pacific, Pacific North America, Central Asia, Europe, Atlantic, Polar, and an European domain specified by Rodwell et al. (2013). Low-skill events were identified meeting certain criteria (events with an ACC less than 0.6 and a RMSE greater than 60 meters, events with an ACC less than 0.5 and a RMSE greater than 60 meters, and events with an ACC less than 0.5 and a RMSE less than 60 meters) in real time which prompted an evaluation of these low-skill events. The evaluation was conducted by first identifying large errors/pattern differences in the five-day 500-hPa geopotential height forecast. These errors were then traced back in time to their original source region. These source regions were then compared with a previous composite study to identify that operational forecasts produce errors through similar processes. For example, low-skill forecasts in the Eastern North America/U.S. sector displayed similar characteristics if a trough propagated across the Rocky Mountains, cutoff lows tried to rejoin the synoptic flow, and/or ridge building occurred in the Pacific.

Title: Representativeness of Coastal Stations for Verifying Open-Water 10 Metre Wind Forecasts

Authors: Nelson Shum[1] and Tim Bullock[1]
1= Environment and Climate Change Canada

When assessing the accuracy of open-water wind forecasts at 10 metres above mean sea level, observations from coastal stations and those located on small islands are often used to augment observations from marine buoys to verify the values predicted. The underlying assumption is that observations from land stations that are well-exposed to the marine environment behave very similarly to observations collected from marine buoys. To test the validity of this assumption, we consider the 10 metre ASCAT (scatterometer) wind fields as a reference; we examine the correlation between wind measurements from coastal stations and the ASCAT open-water wind measurements (in the vicinity of the stations); the same correlation is then made with measurements from marine buoys, and the results are compared. The study shows that despite the proximity to the marine environment of the coastal stations examined, their observed winds show very different characteristics than the winds observed by marine buoys. The results suggest a strong land influence on the coastal station wind measurements, despite the stations' surrounding environment being dominated by water. These findings have implications for how coastal stations should be treated when they are used to verify open-water wind forecasts, and gridded Numerical Weather Prediction forecasts in general.

Title: Object-based verification techniques for short-term thunderstorm forecasts

Authors: Patrick S. Skinner[1], Montgomery L. Flora, Corey K. Potvin, and Anthony E. Reinhart

[1] = Cooperative Institute for Mesoscale Meteorological Studies/National Severe Storms Laboratory - United States

NOAA’s Warn-on-Forecast project aims to produce short-term (0–6 hr) ensemble forecasts of thunderstorm hazards. A prototype Warn-on-Forecast System (WoFS) has produced real-time guidance for select cases since 2016, with forecasts expected to provide accurate guidance of thunderstorms on the scale of a typical National Weather Service warning product, roughly 900 km2. This forecast problem requires a verification framework for discrete events where small location errors are tolerable. As such, an object-based framework has been developed for WoFS evaluation. The object-based techniques used to evaluate WoFS forecasts are based on the Method for Object-based Diagnostic Evaluaton (MODE) and are applied to simulated and Doppler radar-observed proxies for thunderstorms (radar reflectivity) and mesocyclones (updraft helicity and azimuthal wind shear). Forecasts may be evaluated deterministically, by comparing forecast and observed objects, or probabilistically, by identifying coherent regions of overlapping objects as a “probability object”. A single probability of event occurrence is prescribed to these probability objects and matching with observed objects allows for a novel, event-based measure of forecast reliability to be calculated. Additionally, thunderstorm object identification isolates relatively rare features within a forecast domain and facilitates conventional, point-based verification of traditional observations in the near-storm environment. This presentation will provide an overview of the motivation and design of WoFS verification methods and compare object-based results with those from grid-based verification.

Title: A new spatial displacement metric for continuous fields

Author: Gregor Skok[1]
[1] = University of Ljubljana, Slovenia

A new spatial displacement metric called the Neighbourhood Skill Score displacement (dNSS) will be presented. The analysis of idealized and real cases shows that the new metric behaves somewhat similarly to the Fraction Skill Score displacement (dFSS), but with some notable differences. Similarly to dFSS, the dNSS can be used to determine spatial displacement in forecasts in a meaningful way, is not sensitive to noise, with results being directly related to the actual displacements of events, and higher magnitude events having a larger influence on the resulting value. At the same time, contrary to dFSS, the dNSS can also be used when the bias is large, the magnitude of events has a more proportional influence on the results, and, most importantly, the dNSS can also be used for direct analysis of non-binary (e.g. continuous) fields.

Title: Verification of eddy-properties in operational oceanographic analysis systems

Authors: Gregory C. Smith[1] and Anne-Sophie Fortin[1,2]
1 = Meteorological Research Division, Environment and Climate Change Canada
2 = Department of Atmospheric and Oceanic Sciences, McGill University, Montréal, Canada

Recent studies have shown that the presence of oceanic eddies affects the intensification of high-impact tropical cyclones. Many operational weather prediction systems (e.g. in Canada, UK and Europe) have now moved to using fully-coupled atmosphere-ocean prediction models. As a result, the accuracy with which ocean analysis systems are able to constrain the presence and properties of oceanic eddies may affect tropical cyclone forecast skill. While numerous eddy identification and tracking methods have been developed for oceanic eddies, specific methods and metrics tailored to verifying the skill of ocean analyses and forecasts in capturing these features are lacking. Here we apply an open-source eddy-tracking software and adapt it for the purpose of matching eddies between gridded observational analyses and two ocean analysis products of different resolution. A contingency table approach is taken to identify hits, misses and false alarms to provide statistics on the probability of detection and false alarm ratio. These statistics are investigated in terms of their sensitivity to eddy properties (radius, amplitude). Finally, a discussion of the appropriate use of statistical significance in this context is presented.

Title: Improving short-term forecasts of the ocean-sea ice-atmosphere coupled system using wintertime statistics from the MOSAiC campaign

Authors: Jonny Day[1], Amy Solomon[2] and the MOSAiC Near Real-Time Verification Team
1 = ECMWF, 2 = University of Colorado and NOAA/PSL

The MOSAiC Near Real-Time Verification Project (MOSAiC-NRV) is designed to use observations taken during MOSAiC to improve the simulation of coupled processes unique to the Arctic, such as; The vertical representation of cloud and hydrometeors microphysics, low level (mix-phase) clouds; The representation of the stable boundary layer; Atmosphere-snow interaction and ocean-sea ice-atmosphere coupling. Short-term forecasts are used in this project to identify potential errors in the representation of "fast" processes that cause biases in climate model projections of Arctic climate change. The goal of MOSAiC-NRV is to evaluate the skill of fully coupled short-term forecasts during the MOSAiC campaign at the Polarstern location. Multi-model diagnostics focus on process-based evaluation of the coupled system to identify systematic biases that limit the skill of Arctic forecasts. This presentation focuses on the coupled processes that determine the evolution of the surface temperatures and sea ice growth during the winter season (Oct 15 2019-March 15 2020).

Title: Neighborhood-based Continous Ranked Probability Score for Ensemble Prediction Systems

Authors: Joël Stein[1] and Fabien Stoop[1]
1 = MeteoFrance

The neighborhood is widely used by the verification community to reward deterministic forecasts which simulate the right phenomena but not exactly at the right place. For instance, neighborhood-based scores like the Fractions Brier Score is used operationally at Meteo-France to monitor the improvement brought by the high-resolution (HR) AROME model against the low-resolution (LR) ARPEGE model. The neighborhood has been used for ensemble of M forecasts mainly as a post-processing to increase the number of members by considering the M forecasts coming from the neighboring points. It is proposed in this work to include the neighborhood strategy in the CRPS score as it is realized for deterministic forecasts' scores : observations and M forecasts at all points of the neighborhood are used to evaluate the regional distributions of the field and CRPS is then computed from these regional distributions. The fair and unfair neighborhood-based CRPS formulations are presented. Idealized and real cases are presented to demonstrate the advantages of this strategy to reward HR ensembles against LR ensembles. The real cases are Quantitative Precipitation Forecasts (QPF) of the HR ensemble PEAROME and LR ensemble PEARP operational at Meteo-France. The deterministic limit of neighborhood-based CRPS is also presented and used to compare HR deterministic AROME QPF and PEAROME QPF.

Title: Application of neighborhood-based contingency scores to AROME verification

Authors: Fabien Stoop[1] and Joël Stein[1]
1 = MeteoFrance

When verifying high-resolution model forecasts, the double penalty effects affect pointwise scores and can erroneously lead to the assumption that global models perform better than high-resolution models. A way to circumvent these effects is the use of neighborhoods, such as in the Fractions Brier Score (FBS) computation. In particular, Météo-France uses a derived version of the FBS, the Frequency Brier Skill Score against the persistance forecast (FBSS), as headline score to evaluate the quality of it high resolution model AROME. However, several user feedbacks indicate that the FBS and FBSS are difficult to interpret, and that they prefer the use of contingency scores such as Hit Rate, False Alarm Ratio or Peirce Skill Score.

A new method to introduce neighborhoods directly in contingency tables has been developed at Météo-France (Stein and Stoop 2019) : in a local neighborhood misses and false alarms compensate each other. The quantitative precipitation, wind gust and visibility forecasted by ARPEGE (Météo-France global model) and AROME are verified through these neighborhood-based contingency scores. It permits to compare the quality of ARPEGE and AROME models by limiting the double penalty effects through the use of contingency scores easily understandable by users. In particular, classical skill scores based on these neighborhood contingency tables (Pierce skill score, Heidke skill score,...) will be used to replace the current AROME headline score.

Title: Huber loss as a scoring function for forecast verification

Author : Rob Taggart[1]
1 = Australian Bureau of Meteorology

If an organisation requests its forecasters to issue single-valued forecasts but doesn't give guidance on which point (e.g. mean) of a predictive distribution should be quoted, then how should one retrospectively assess the quality of such forecasts? This dilemma was faced by a team within the Australian Bureau of Meteorology. One method of assessment uses a scoring function to rank forecasts. Two possible candidates – absolute error and squared error – each have downsides. Absolute error does not distinguish between the two error sequences (4, 0, 0, 0) and (1, 1, 1, 1), and it was argued that perhaps four smaller errors should be preferred over one larger error. On the other hand, squared error is sensitive to large errors, which is problematic if a large error is caused by an undetected faulty observation rather than a poor forecast. Instead, a compromise scoring function was used: Huber loss. Huber loss applies a squared loss penalty for small errors and a linear penalty for large errors. It has been used in robust regression for decades. This talk examines the use of Huber loss as a scoring function. In this context, a forecaster who wants to optimise their score should quote the "Huber functional" of their predictive distribution. Quantiles (e.g median) and expectiles (e.g. mean) are limiting cases of the Huber functional, as is reflected by new theorems characterising its consistent scoring functions and their mixture representations. The use of Huber loss for robust verification of expected value forecasts will also be discussed.

Title: Representation of process based diagnostics in NCUM global and regional models

Authors: Mohana. S. Thota[1], Kondapalli Niranjan Kumar, Sagili Karuna Sagar, Raghavendra Ashrit
1 = National Center for Medium Range Weather Forecasting, India

Primary goal of this study is to assess the fidelity of National Center for Medium Range Weather Forecasting Unified model's (NCUM) global (12km) and regional (4km) versions in representing the monsoon sub-seasonal variability over Indian region by applying process based diagnostics. Moisture budget analysis is performed on the model's forecast fields for a typical extended monsoon episode event occurred during boreal summer monsoon season 2019. The exercise is repeated using the ERA5 reanalysis and the relative roles of the budget terms are quantified. We also tested the budget diagnostics onto the newly generated Indian Monsoon Data Assimilation and Analysis (IMDAA) product. Preliminary results obtained form the moisture budget analysis is encouraging. Analysis indicate that, dry air advection from the northwest regions strongly dries the atmospheric column nearly 7-10 days before the peak dry phase over Indian subcontinent. Dry air intrusion towards Indian subcontinent is consistent with the anomalous total precipitable water vapor from satellite observations. One of the main implication of this work is, the lead time obtained in the moisture advection term can be used to improve the model forecasts.

Title: Developing a Space Weather Verification System Using METplus

Authors: Jonathan Vigh[1], Terrance Onsager, Robert Steenburgh, Tara Jensen, Barb Brown, John Halley Gotway, Tatiana Burek, George McCabe
1 = National Center for Atmospheric Research, USA

Model Evaluation Tools (MET) is a highly-configurable, state-of-the-art suite of verification tools developed by the National Center for Atmospheric Research (NCAR) and supported to the Development Testbed Center. The extension of MET to METplus wraps these tools with python, allowing for complex verification workflows to be simplified and adding generalized capability to read any model data source with user-written python codes. While METplus was originally developed for verification and evaluation of atmospheric numerical weather prediction models, it is being continually developed and expanded into new prediction domains, such as subseasonal-to-seasonal forecasts, aerosols and air quality, hurricanes, and upper atmosphere. Through a collaboration between NCAR and NOAA's Space Weather Prediction Center (SWPC) METplus is being expanded into supporting a host of space weather verification and diagnostics. The system will extend beyond the current Ionospheric Total Electron Content (TEC) work already completed. Ultimately, the system will provide statistics on SWPC's human-generated forecast products, as well as evaluation of models (e.g., geospace, solar wind, geoelectric), and comparison and diagnostics of model differences. This conference paper summarizes progress toward the goal of developing a real-time verification system for SWPC products and provides initial evaluation results.

Title: Process-oriented Model Diagnostics for Extended-range Forecasts

Authors: Zhuo Wang[1], Jiacheng Ye, Tara Jensen, Doug Miller, Weiwei Li
1 = Illinois University

Model validation and evaluation is an indispensable part of model improvement efforts. While performance-oriented metrics provide quantitative measures on how well a model does, process-oriented metrics help to reveal model deficiencies and identify pathways to model improvement.

A suite of process-oriented, observation-based model diagnostics are developed to evaluate the processes that are critical to forecasting on the synoptic to subseasonal time scales. The suite consists of three levels of diagnostics: i) evaluation of systematic model errors in representing moist convection and cloud processes; ii) evaluation of the sources of predictability relevant on S2S timescales (such as the MJO, NAO and weather regimes); iii) evaluation of high-impact weather systems (such as tropical cyclones, blocking, etc.).

The presentation will illustrate examples for each level of diagnostics using the GEFS retrospective forecasts. The diagnostics will be made available to the community via the Model Evaluation Tools (METplus) and the Model Diagnostics Task Force (MDTF) Diagnostic Package.

Title (keynote): Validation and Verification of Climate Products

Authors: E. C. Weatherhead[1], N. Georgas, A. F. Blumberg
1 = Colorado University, Boulder, CO, USA

Increasingly, governments, corporations and individuals are using climate information to make decisions about investments, resources and property. These new products often help with risk assessment for a variety of perils: drought, flooding, fires, severe storms, extreme heat as well as highly specialized products tailored to specific health risks, individual crops or environmental vulnerabilities. In many cases, estimates for these perils are provided for thirty years into the future or longer. For all of these products, important questions to ask are “How valid are the methods used to create these products?” and, often of higher relevance, “How can such products be verified?” Climate models continue to be our best tools for understanding future climate; estimating their skill for common perils is challenging because most of the events of concern are extreme events. Developing valid approaches requires using advanced extreme event statistics to assure appropriate use of climate models. Verifying the results of these approaches requires comparison to all available information, including historical events, heuristic models, and current conditions. Four specific approaches can be used to verify climate products: comparison to current spatial patterns and information; comparison to current temporal distributions; evaluation relative to other climate products; examination of specific extreme events. This presentation will outline both the validation and verification approaches in use for verification products, and show how these metrics can be used to quantify quality of the products for decision support.

Title: Verification of a prototype wind impact forecast using building damage reports

Authors: David Wilke[1], Harald Richter, Elizabeth Ebert, Craig Arthur, Mark Dunford, Martin Wehner
1 = Bureau of Meteorology, Australia

In recent years the international community has moved toward the provision of impact and impact-based forecasts in order to enhance the utility of traditional weather forecasts. Verifying such forecasts, however, presents considerable challenges. Here we assess the performance of a prototype wind impact forecast with respect to individual residential building damage reports obtained during a high-impact weather event that devastated parts of the state of New South Wales, Australia in 2015.

The impact 'forecast' is produced using available exposure, vulnerability and 1.5 km reanalysis hazard data and assigns a mean damage state, categorised between 1-5, to geographic areas with a population of roughly 200-800 persons. We compare to observational data which records structural damage to individual assets in 5 categories in addition to a range of other fields, including, for example, the presence of any water inundation. The observations are filtered to remove reports that are unlikely to be a result of wind damage. They are also aggregated over the geographic areas for comparison at the forecast grid scale. For reference, the forecast is contrasted with a straight-forward implementation of static wind-warning criteria employed by the Australian Bureau of Meteorology.

Our analysis shows that while the performance in this case is modest, the new impact forecast out-performs the reference forecast. The verification procedure also highlights the value additional observational fields, particularly regarding hazard-damage linkages, could provide in estimating future impact forecast performance.

Title: Verification of subseasonal sea-ice prediction at both poles

Authors: Lorenzo Zampieri[1], Helge F. Goessling, Thomas Jung
1 = Alfred Wegner Institute, Germany

With retreating sea ice and increasing human activities comes a growing need for reliable sea-ice forecasts up to months ahead. We exploit the subseasonal-to-seasonal (S2S) prediction database and provide a thorough assessment of the skill of operational forecast systems in predicting the location of the Arctic and Antarctic sea-ice edges on these time scales. This study employs the Spatial Probability Score, a probabilistic verification metric specifically designed to capture the correctness of the sea ice edge position. Our verification methodology goes beyond the classical sea-ice extent and area, and tries to provide a sea-ice forecast description that could be valuable for planning shipping operations. We find large differences in skill between the systems, with some showing a lack of predictive skill even at short weather time scales, and the best producing skillful Arctic forecasts more than 1.5 months ahead. We assess the forecast skill in both hemispheres, thereby showing that prospects for subseasonal sea-ice predictions are promising, especially for Arctic late summer forecasts. To fully exploit this potential, it will be imperative to reduce systematic model errors and develop advanced data assimilation capacity. Furthermore, the relatively long time-span of the S2S prediction database – which is more than 20 years for some of the considered forecast systems – allows us to present some considerations about the changes in predictive skills as the sea-ice extent and volume decreases.