

Verification of Environmental Prediction in Polar Regions: Recommendations for the Year of Polar Prediction

B. Casati, T. Haiden, B. Brown, P. Nurmi, J.-F. Lemieux

1. Motivations and Aims

Recent years have seen an increasing interest in the performance of environmental prediction systems in polar regions, driven mainly by three factors: i) the Arctic amplification of climate change signal, and the risks and opportunities associated with anthropogenic induced global warming; ii) increasing human activities in these regions, such as land and marine transportation, tourism, natural resources exploitation, fishing and other related economic activities; and iii) the linkages, interactions and impacts that polar weather has on mid-latitudes.

The main goal of the Year of Polar Prediction (YOPP) is to improve understanding and the representation of key polar processes in numerical weather and climate prediction models. Given the dependence between polar and mid-latitude weather, one of the YOPP secondary outcomes is to enhance large-scale predictive skill beyond polar regions. To achieve these goals, YOPP is coordinating an extensive period of intensive observing and modeling activities (mid 2017 - mid 2019), along with verification, user-engagement and educational activities.

The identified major verification goals within YOPP are to: i) obtain quantitative knowledge of model performance (both for model developments and for user-relevant applications); and ii) compare (different systems) and monitor progress with respect to the present-day base-line model performance. These seemingly simple goals in reality encompass a large spectrum of verification tasks.

Some key factors need to be considered in order to implement and apply a successful verification strategy. Specifically, the following questions must be answered:

- i. Who are the verification end-users (e.g., modelers or forecast end-users, such as navigation companies)?
- ii. What are the verification purposes (e.g., diagnostics or administrative)?
- iii. What questions need to be addressed (e.g., model predictability limit) and/or what are the forecast attributes of interest (e.g., timing for onset and clearance of fog)?
- iv. What type of forecast is to be verified (e.g., deterministic -continuous or categorical- or probabilistic)?
- v. What are the statistical characteristics of the variables to be verified (e.g., smooth upper-air variables, such as geopotential height and temperatures, or spatially-episodic and discontinuous variables, such as precipitation or sea-ice)?
- vi. What are the available matching observations (e.g., *in-situ* measurements or satellite-based spatial observations)?

The first three points aim to properly formulate the verification questions, whereas the last three points address some of the technical aspects of the verification strategy to be implemented. No single

verification technique can possibly address all verification users/purposes/questions and/or be suitable for all forecasts/variables/observation types; each verification strategy should be tailored to the users' needs and their verification purposes and questions, and on the forecasts and variables verified, as well as the corresponding available observations. In this report we aim to provide some recommendations for verification of environmental prediction systems in polar regions, with the choice of recommended verification strategies based on the aforementioned factors.

1.1. Verification purposes and end-users

Prediction, and therefore also verification, within YOPP addresses a wide and heterogeneous community of users. On one side there are Numerical Weather Prediction (NWP) developers, who need informative diagnostics for better understanding the capability and/or shortcomings of current numerical models in reproducing key polar physical processes. Summary verification scores are used by weather services, model developers, researchers and generic users for i) assessing NWP forecast quality, including skill (e.g., versus persistence or climatology), ii) exploring the predictability limits of current NWP systems, iii) comparing different NWP systems (or different configurations and physics), iv) monitoring progress (e.g., comparing pre-YOPP and post-YOPP performance), v) comparing forecast performance in the mid-latitudes versus polar regions. Summary verification scores can also be exploited by sophisticated forecast end-users through application of cost/loss scenarios to measure the value added by NWP forecasts in polar regions. Finally, forecast end-users such as commercial business and government departments (e.g., marine transportation, aviation) might benefit from easy-to-interpret physically meaningful verification metrics specifically designed to respond to their needs, to help planning of their activities. This list is not exhaustive, but already encompasses a large variety of user needs; in particular, a variety of variables and different forecast time-scales and lead-times, as well as different verification approaches must be adopted to address the different needs of the variety of forecast users.

In this report, recommendations for the verification strategy to be adopted within YOPP are made with respect to three classes of users and verification purposes: diagnostics for model developers (Section 3.1); summary verification scores for administrative and generic purposes (Section 3.2); and physically-meaningful verification measures for forecast end-users (Section 3.3). This latter class is vast (and could result in a wide variety of approaches); therefore in this report we focus only on verification of sea-ice predictions to illustrate the potential usefulness of some physically meaningful spatial verification approaches on one exemplar single variable.

1.2. Variables and key polar processes

Within the YOPP implementation plan, the following key variables and physical processes have been identified:

- Basic (surface and upper-air) atmospheric variables: temperature and dew-point temperature, precipitation, cloud cover, relative humidity, wind (speed and direction), geopotential height, mean sea level pressure.
- Environmental surface variables: sea-ice, snow at the surface (snow cover, snow thickness),

permafrost (soil temperature).

- Modeling challenges (processes/variables):
 - Coupling between atmosphere - land - ocean - cryosphere: in polar regions, verification could focus on the (presence/absence of) snow / sea-ice, and their effects on the interactions (e.g., flux exchanges) between land / ocean and atmosphere, as represented by numerical models;
 - Surface-atmosphere exchanges: these include the validation of the radiative transfer and the energy, moisture and momentum fluxes. As an example, variables of interest could be latent and sensible heat fluxes, which are related to the humidity and temperature vertical transport between the surface and the atmosphere.
 - Stable boundary layer representation: this involves the analysis of boundary layer turbulence, and temperature and wind vertical profiles in stable regimes.
 - Effects of steep orography (e.g. orographically enhanced precipitation);
 - The representation of clouds, with specific focus on low-level mixed-phase clouds.
- High Impact Weather: polar lows, low-level jets, topographically influenced flows such as katabatic winds and hydraulic shocks, extreme thermal contrasts, blizzards, freezing rain, fog.
- User-relevant variables: Visibility, ceiling and icing (for aviation); sea-ice, fog and visibility (for navigation); ground conditions (e.g. snow, permafrost) for land transport.

Different users might be interested not only in verifying different variables, but also in evaluating different aspects of the same variable. As an example, NWP model developers might be interested in sea-ice concentration (presence or absence) and sea-ice thickness, because of the sea-ice effects on the radiation budget (e.g., albedo) and surface atmospheric variables (e.g., surface air temperature), and because of the role sea-ice plays as a buffer between the ocean and the atmosphere (i.e., affecting energy fluxes and ocean-atmosphere coupling). Sea-ice pressure, on the other hand, is extremely important for shipping, for defining a navigation route. The specific variables/processes of interest for each of the users identified in Section 1.1 are listed in Section 3, along with recommended verification approaches.

2. Observation challenges

Observations are the cornerstone of verification. However, polar regions are characterized by harsh environmental conditions and are hardly populated; hence surface-based observations are difficult to obtain and scarce. One of the greatest challenges of verification in polar regions is the limited amount of reliable observations.

Surface observations are sparse and observation networks are usually not homogeneous across the domain (i.e., stations are often unevenly distributed in space, with observation network being more dense in more populated regions). In addition, observation sites are often not fully representative of the whole polar environment (e.g., most of the stations are located along the coast). Surface observations are also characterized by a strong seasonality, with fewer records in winter and more in the summer.

Moreover, instrument failure can affect observations in the harsh polar environment, and observation uncertainty related to these faulty measurements is difficult to detect due to lack of nearby buddy-check stations.

To date, only a few verification studies have addressed the issue of observation uncertainties and space-time representativeness of observations (or sampling uncertainty in space and time). Ciach and Krakewski (1999) proposed approaches for coping with observation errors in computation of root-mean-squared error (RMSE) values. Bowler (2008) considered how to incorporate observation uncertainty into categorical scores, and Santos and Ghelli (2011) propose a variation of the BSS that accounts for observation uncertainty. Ahrens and Jaun (2007) verified ensemble forecasts against ensembles of analyses obtained by stochastic interpolation of point observations, using the Brier Skill Score (BSS). Saeltra et al (2004) analyze the effects of observation errors on rank histograms and reliability diagrams. Candille et al (2007) evaluate the ensemble dispersion while accounting for the observation uncertainty. Mittermaier (2014) explored the impact of temporal sampling on the representativeness of hourly synoptic surface temperature observations. Casati et al (2014) propose a spatial wavelet-based verification approach which accounts for the inhomogeneous spatial density of station observation networks across a domain. Some spatial neighborhood verification approaches (e.g., Theis et al, 2005; Atger, 2001) can be applied to station observations, and the time dimension might compensate for station spatial sparseness. However, most of these approaches are still experimental and not used routinely in NWP verification. *YOPP could serve as a test-bed for new research into verification approaches that account for observation sparseness and uncertainty (with a focus on the polar context), which would benefit NWP verification in general.*

Given the sparseness of surface observations in polar regions, YOPP will need to exploit satellite-based observations. A great advantage of satellite space-borne products is not only the enhancement of the observation spatial coverage, but also the availability of observations that are spatially defined, which enable, for example, the detection and comparison of spatial patterns. The availability of spatial observations opens the possibility of using modern spatial verification techniques (Gilleland et al 2010). Spatial verification techniques account for the coherent spatial structure (and the presence of features) characterizing weather fields, and these approaches provide more physically meaningful and diagnostic results than traditional verification approaches. A concise review of these techniques is given in Appendix A.

Requirements for satellite space-borne atmospheric observations to be used for evaluation of polar predictions include a good representation of lower atmospheric structure (e.g., high-resolution wind, temperature, moisture profiles), clouds (e.g., liquid versus ice phase profiles, particle size distributions, aerosol concentration and type) and snow-cover (depth, layering, snow water equivalent, melting ponds, albedo, temperature). However, the use of visible and infra-red (IR) satellite observations for characterizing the atmosphere in polar areas is currently limited, mostly because the lower troposphere is nearly isothermal and often cloud covered, and the optical properties of snow/sea-ice covered surfaces are difficult to characterize; these factors clearly limit the use and effectiveness of temperature and moisture sounder data. Verification activities in polar regions can significantly benefit from advancements in satellite technology and the use of several and diverse space-borne instruments (e.g. visible and IR can be complemented by passive microwave sounders). Verification techniques should account for the challenges and limiting factors of satellite retrievals in polar regions, for example, by

including observation uncertainty in their scoring algorithm. Uncertainty in satellite-based observations can be quantified by using multiple observation products retrieved from different satellites (e.g., temperature and humidity can be retrieved from AMSU-A, AMSU-B and ATMS imageries onboard different satellites). Finally, atmospheric variables (e.g., temperature and humidity) retrieved from satellite observed radiances are synthesized based on physical and statistical remote-sensing assumptions, which possibly affect the verification results. To mitigate the effects of these assumptions (and their associated uncertainties), verification can be performed with a model-to-observation approach, using, for example, model-simulated brightness temperatures for a direct comparison against satellite-retrieved brightness temperatures.

Data assimilation algorithms are often used to harmonize and merge radiances from different satellites (as well as observations from different sources). These algorithms perform quality controls and bias corrections which can be influenced by the background state. This procedure introduces an undesired dependence between verifying observations and the model itself, which needs to be taken into account in the interpretation of the verification results (the analysis-model dependence is further discussed later in this section). On the other hand, data assimilation algorithms rely on information on the model and observation errors, and could possibly be exploited for providing an estimate of the observation uncertainty. However, this should be investigated with caution: in fact, data assimilation performs several assumptions about the structure and dependencies of the background model and observation errors (Hollingsworth and Lönnerberg, 1987), and the observation uncertainty is often inflated to optimize its use for data assimilation purposes (Liu and Rabier, 2003); therefore, the observation uncertainty used in the data assimilation process might not be optimal for verification purposes, but data assimilation statistics might still be informative, when non-inflated estimates are considered (Desroziers et al, 2005).

Satellite-based sea ice products are crucial for navigation at high latitudes. In fact, space-borne measurements can determine sea-ice concentration, thickness, and the location of icebergs. Moreover, satellite-based sea-ice tracking systems (e.g., Komarov and Barber, 2014; Figure 7) can provide information on the sea-ice drift and deformation (spatial gradients of sea-ice velocities). Some current operational sea-ice prediction systems have sufficiently high spatial resolution (5 to 1 km), which allows them to simulate features such as leads and pressure ridges. However, the resolution of certain technologies of satellite imagery is still coarse (~50km for SSMIS passive microwave sounders), and these small-scale phenomena are not visible in their associated products. As a consequence, high-resolution sea-ice models can be penalized in verification practices for producing these non-observed small-scale features. Visible and infra-red satellite products, e.g. from Advanced Very High Resolution Radiometers (AVHRR), can attain finer resolutions (up to 1km) and observe the small scales of such user-relevant phenomena. However visible and infra-red satellite imageries are still affected by the lack of contrast between cloud cover and sea-ice. Synthetic Aperture Radar (SAR) imagery such as Sentinel-1 and RADARSAT-2 are needed in order to match the high resolution of sea-ice models and correctly characterize sea-ice versus cloud cover (e.g. see Figure 1). Several advanced satellite-based sea-ice gridded products are already available from national ice services: these products include the ice charts produced by the Canadian Ice Service (<http://ice-glaces.ec.gc.ca>) and the sea-ice products produced by the National Oceanic and Atmospheric Administration (NOAA) Ice Mapping System (IMS; <http://www.natice.noaa.gov/ims/index.html>), which should be considered for sea-ice verification in YOPP.

Verification against gridded datasets (and/or analyses) has two major advantages: i) the observation quality control, ingestion of the observation uncertainty, and representativeness issue are dealt with in the gridding process; and ii) the observation field is spatially defined (and it covers the whole space-time domain). The latter advantage makes it possible to implement spatial verification approaches (see Appendix A) and opens more interesting options for informative graphical display of verification statistics, such as Hovmoller diagrams, zonally and meridionally averaged scores versus the lead time or a vertical profile, and so on (e.g. Figure 3). However, within the gridding process several assumptions are introduced; for example, the use of a kriging process to fill-in the space between point observations requires assumptions regarding the representativeness of the observations ingested. Verification against gridded datasets (and/or analyses) must be performed with awareness of the strengths and weaknesses of the specific gridded dataset used.

Verification of a model-based forecast against its own model-based analysis is affected by their inter-dependence (e.g. Figure 4), and it is therefore essential to acknowledge the caveats and drawbacks associated with this verification practice. As an example, Park et al (2008) compare the performance of eight ensemble prediction systems from the TIGGE archive against analyses, and show that verification of each EPS against their own analysis leads always to the best score: thus, caution in the interpretation of the verification results must be used when ranking different numerical prediction systems by verifying them against a single model-based analysis. Similarly, decisions on the developments of a numerical model should not be based (solely) on verification results against its own analysis, since this might lead to drifting away from reality. *Verification studies that assess the impact of using model-dependent analyses (versus observations) are sought (such impacts are expected to be larger in polar regions than in mid-latitudes, due to the limited numbers of observations).*

Model biases in polar regions are large compared to the biases in mid-latitude regions, and data assimilation systems are sub-optimally adapted to polar conditions; thus, many observations are rejected or given inappropriate weight. As a result, model-based analyses in polar regions might lean towards their background model, more than for mid-latitudes. Bauer et al (2014) compare five analyses from the TIGGE multi-model ensemble in the Arctic: they found that the spread between the TIGGE multi-model analyses exhibits much larger discrepancies with respect to the analysis uncertainty estimated by a single-model ensemble data assimilation system. They conclude that neither current multi-model analyses (possible over-dispersive) nor ensemble data assimilation (possibly under-dispersive) properly represent polar analysis uncertainties. *YOPP could serve as a platform for enhancing synergies between the verification and data assimilation communities: verification could better inform data assimilation about model biases, observation errors, spatial and temporal representativeness issues in polar regions; data assimilation could use that information in their error models and, in return, provide a vast number of observation-model statistics from the assimilated data. Both communities could gain from shared knowledge on representativeness and observation uncertainties, and shared tools for (model-independent) quality controls.*

A good verification practice is to perform verification solely against analysis values which are based on recently assimilated observations (as opposed to model-based values): as an example, Lemieux et al (2016) performed a model-to-analysis verification which solely considered the analysis grid-points where the latest-assimilated satellite-based observation is more recent than 12 hours. Verification

against multiple gridded datasets and/or analyses is recommended: the uncertainty/spread between analyses/gridded observation datasets should be an order of magnitude smaller than the forecast error. The assessment of models versus analysis uncertainties and errors can be accomplished by using multiple models (e.g., TIGGE, YOTC, Transpose AMIP) and multiple re-analyses (e.g., ERA-Interim/20C, JRA-55, MERRA-2, Arctic System Reanalysis, Climate Forecast System Reanalysis). Some verification scores and statistics (e.g., Brier Score, CRPS, KS distance), can directly compare the distributions derived from an ensemble prediction system (or an ensemble of different models) and an ensemble of analyses.

Different challenges are associated with each observed variable because the verification observations associated with each variable are obtained from measurements with different characteristics, with different uncertainties and that are synthesized based on different assumptions. The strengths and weakness of each variable and (gridded) observation dataset should be known: accomplishing this is challenging since it encompasses expertise from many different fields. Where possible, YOPP verification tasks should be repartitioned (especially for specific user-relevant variables) to represent the interests of each involved agency/stakeholder.

3. Verification approaches

3.1 Model diagnostics

Model diagnostic verification aims to assess specific model behaviours, for a better understanding and improvement of key physical processes and their representation in numerical modeling. Such process-based diagnostic verification is used to compare different NWP physical schemes and parameterizations. The end-users of such process-based diagnostic verification are de-facto the model developers.

Process-based diagnostic verification usually assesses all the physical aspects of a few targeted and well-observed cases studies. These case studies are often identified within an intensive observing period with high resolution and high frequency observations. Process-based diagnostic verification within YOPP could be performed at super-sites, which comprises multi-variate observations with high temporal resolution. The Arctic super-sites identified in the YOPP implementation plan include Sodankylä (FMI Arctic research centre, <http://fmiaarc.fmi.fi>); Svalbard Integrated Observing System (SIOS <http://www.sios-svalbard.org/>); International Arctic Systems for Observing the Atmosphere (IASOA, www.iasoa.org) stations such as Tiksi, Summit, Eureka, Alert, Barrow; and the Russian drifting North Pole station. In the Antarctic the super-sites include Dome-Concordia and South Pole. Observations from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAIC) campaign could be valuable for a post-YOPP evaluation of improved NWP systems' capabilities.

Model diagnostics often address the verification of specific physical processes, where model outputs are compared with observations of process-specific physical variables (e.g. latent and sensible heat fluxes). For example, the GABLS-4 project (<http://www.cnrm.meteo.fr/aladin/meshtml/GABLS4/GABLS4.html>) undertook an inter-comparison of the capabilities of several single-column, land-surface, and large-eddy simulation models to represent a

strongly stable boundary layer in Antarctica: model evaluation focused on turbulent fluxes of temperature, humidity and momentum. *Process-based model diagnostics are very specific and ideally should be undertaken by (or outlined in close collaboration with) model developers.*

Some of the key model processes in polar regions, and associated physical variables, have already been listed in Section 1.2. Note that evaluation of these processes involves verification of process-specific physical quantities (e.g., energy, moisture and momentum fluxes; radiation budget), beyond traditional surface and upper-air physical variables. These model-specific variables are sometimes not directly observed; in these cases, model behavior is usually assessed based on theoretical expected outcomes, or against analyses (which also include model-specific variables). Model diagnostics against analyses can be informative and can be practiced, as long as caveats with respect to verification against model-based analyses and gridded observation products are known and accounted for (see discussion in Section 2). In general, it is recommended that assessment of different model configurations and parameterizations should be based on comparisons to actual observed values, or by using model-simulated retrieved variables (e.g., brightness temperature) to more directly evaluate measured phenomena.

In current practice, model diagnostics favor the use of simple yet informative summary statistics (e.g., the additive bias) graphically displayed along the vertical profile (e.g. Figures 4 and 5) and/or for the diurnal cycle and/or zonal averages (e.g., Figure 3, left panel). A meaningful graphical display, in this context, is fundamental; for example, a Hovmoller diagram can help detect flow-dependent error propagation (Figure 3, right panel). Direct visual (eye-ball) verification of the observed and modeled physical variables/phenomena of interest is often the most effective approach.

Current model diagnostics tend to still focus on individual parameters. However, given the importance of better understanding the entire physical process (characterized by strong interactions between the different physical variables), model diagnostics could be more informative through use of multi-variate statistics.

Some of the recently developed spatial verification approaches (see appendix A) could be useful for diagnosing model processes. As an example, field-deformation and feature-based approaches could provide feedback on model advection. High resolution models could benefit from neighborhood verification approaches. Timing errors associated with the offset of specific physical processes and/or weather phenomena can be assessed by the use of lag correlation of time series, possibly within a neighborhood verification approach.

Aggregation across multiple cases (or for a season) can provide more robust statistics; however, a targeted stratification (or conditional verification) can provide more informative verification results. As an example, conditional verification performed on multiple variables targeting a specific process (e.g., liquid precipitation in the presence of a temperature inversion, whereby temperature near the ground is below freezing, for freezing rain) can help diagnose process-related model deficiencies. Model developers are encouraged to optimize the delicate balance between aggregation and stratification (i.e., conditional verification). Inference in model diagnostics is essential, given that model diagnostics typically aim to compare the effects of different model configurations and parameterizations. Inference is briefly discussed in the next session.

3.2 Summary verification scores

Verification of the basic surface and upper-air atmospheric variables should be used for monitoring and comparing NWP systems. As minimum standard, YOPP should aim to meet the WMO Commission for Basic Systems (CBS) recommendations (WMO-485, WMO-893, Haiden et al 2014, WMO-1091) summarized in Tables 1a,b. CBS mandatory and additional recommended surface and upper-air variables are: 2m temperature, 2m dew-point temperature, 2m relative humidity; 24h and 6 h accumulated precipitation; 10m wind speed and direction; total cloud cover; mean sea-level pressure (mslp); relative humidity, wind components and geopotential heights at different vertical levels (850, 500, 250, 100 hPa). Additional vertical levels could be considered (e.g. to sample the stratosphere). Moreover, the CBS recommends to compute scores for forecasts initiated at 00:00 and 12:00 UTC separately, with a frequency of 12 hours for upper-air variables, and a frequency of 6 hours (3 hours up to T+72 hour forecast lead-time) for surface variables. Note that these recommendations are based on the minimum standards, documented in the Manual of the Global Data-Processing and Forecasting System (WMO-485), for availability of NWP fields by NWP producing centers, and are constrained by observation frequencies. *If YOPP benefits from more frequent surface and/or upper-air observations, verification initial times and frequencies should adapt accordingly (e.g., higher verification frequency could better detect signals related to the diurnal cycle).*

Table 1a: Simplified summary of CBS standards (pending confirmation by CBS in 2016 for surface variables) for verification of deterministic NWP products. Upper-air variables and levels shown for extra-tropics only. Listed are mandatory requirements, with additionally recommended items in parentheses.

	Upper air	Surface
Variables	Mean sea-level pressure, geopotential height, temperature, wind (relative humidity)	2m temperature, 10m wind speed and direction, 24h precipitation (total cloud cover, 6h precipitation, 2m relative humidity, 2m dew-point)
Levels (hPa)	850, 500, 250 (100)	
Frequency	24 h (12 h)	6 h up to T+72h, 12 h afterwards (3 h up to T+72h, 6 h afterwards)
Scores	Mean error, root mean square error, anomaly correlation, S1 score for mslp (mean absolute error, rms forecast and analysis anomalies, standard deviation of forecast and analysis field)	Mean error, mean absolute error, root mean square error, contingency tables [see thresholds below]
Thresholds for contingency tables		10-m wind speed: 5, 10, and 15 m s ⁻¹ 24-h precipitation: 1, 10, and 50 mm 6-h precipitation: 1, 5, and 25 mm Total cloud cover: 2 okta, 7 okta
Interpolation	Nearest grid-point on native model grid; interpolation to 1.5x1.5 deg grid for verification against analysis	Nearest grid-point on native model grid

Table 1b: Simplified summary of CBS standards (pending confirmation by CBS in 2016) for verification of probabilistic NWP products.

Variables	Mean sea-level pressure, 500 hPa geopotential height, 850 hPa temperature, 850 hPa wind speed, 850 and 250 hPa wind components, 24h precipitation
Frequency	24 h
Scores	Continuous ranked probability score (CRPS) Brier Skill Score (with respect to climatology) Relative Operating Characteristic (ROC) Relative economic value (C/L) diagrams Reliability diagrams with frequency distribution Spread (standard deviation of ensemble)
Thresholds for contingency tables	PMSL anomalies: ± 1 , ± 1.5 , ± 2 standard deviations 500 hPa geopotential height anomalies: ± 1 , ± 1.5 , ± 2 standard deviations 850 hPa wind speed: 10, 15, 25 m s ⁻¹ 850 and 250 hPa u and v wind components: 10th, 25th, 75th and 90th percentiles 850 hPa temperature anomalies: ± 1 , ± 1.5 , ± 2 standard deviations 24h precipitation: 1, 5, 10, and 25 mm
Interpolation	Nearest grid-point on native model grid; interpolation to 1.5x1.5 deg grid for verification against analysis

Traditional summary measures of performance and skill are recommended.

- Continuous scores are recommended for deterministic forecasts of continuous, normally distributed and spatially smooth variables (e.g. temperature, sea-level pressure, geopotential height). Continuous scores include bias, MSE, MSE Skill Score (versus persistence and climatology), MAE, (anomaly) correlation, S1 score. The performance of several NWP systems can be compared by displaying their continuous scores on Taylor (2001) diagrams.
- Categorical verification is recommended for deterministic forecasts of right-skewed, episodic or spatially discontinuous variables (e.g. precipitation, wind, clouds). Categories are defined by user-relevant thresholds, and then categorical scores are evaluated from the contingency table entries. These includes: FBI, TS and ETS, PC, HSS, OR, YQ. Summary performance diagrams (e.g. Roebber, 2009) can be used to display several categorical scores and compare different models.
- Traditional continuous and categorical verification scores degenerate to un-informative trivial values as events becomes rarer (Stephenson et al 2008). *Extremes dependence indices (EDS, EDI, SED, SEDI) are recommended for the verification of extreme and rare events (Ferro and Stephenson, 2011).*
- Recommended verification approaches for ensembles and probabilistic forecasts include: the Brier Score and Brier Skill Score (and their resolution+reliability+uncertainty decomposition), the CRPS (and its resolution+reliability decomposition), ROC and reliability diagrams, Rank histograms and the dispersion score obtained from the Reduced Centered Random Variable (RCRV; Talagrand et al, 1999; Candille et al, 2007), discrimination diagrams and the Generalized Discrimination Score (Weigel and Mason, 2011), and the analysis of the ensemble error-spread relationship (Christensen et al., 2014). Several examples of ensemble verification can be found on the TIGGE museum web-page (<http://gpvjma.ccs.hpcc.jp/TIGGE>), developed

and maintained by Prof M. Matsueda, and in Jung and Matsueda (2014), Jung and Leutbecher (2007).

Summary verification scores can provide information on several aspects of the model performance and serve several purposes (beyond monitoring and comparing forecasting systems). Traditional skill scores can assess model performance versus persistence and climatology, and investigate the predictability limits of present NWP systems in Polar Regions: for example, predictability (e.g., in terms of forecast lead-time) might be a base-line to be beaten within the YOPP modeling effort. Predictability as a function of forecast origin time can reveal key processes and variables which (when assimilated) lead to significant improvements in polar prediction capabilities: this kind of information could be relevant for both model developers and some forecast end-users (e.g., Day et al, 2014). As an example, early season sea-ice extent affects the length of the navigation season: since sea-ice seasonal forecasts initiated in June have enhanced predictive power, they should be prioritized by shipping companies in their planning. Traditional ensemble diagnostics can assess the consistency of the different NWP forecasts (e.g., the ensemble error-spread relationship), and possibly can help identify the regions / processes / variables characterized by large spread (i.e., model uncertainty).

A primary goal for verification within YOPP is to identify the sources of systematic forecast errors; while aggregation is fundamental in order to obtain useful and potentially significant verification results, an optimal and tailored stratification can be crucial for revealing process-related or flow-dependent errors. *Conditional verification* performed on specific weather regimes (e.g., Crocker et al, 2014) or on multiple variables targeting a specific process (e.g., liquid precipitation in the presence of a temperature inversion, whereby temperature near ground is below freezing, for freezing rain) can help diagnose some flow-dependent systematic errors and/or process-related model deficiencies.

Spatial verification approaches (see Appendix A for a concise review) can also help characterize the origin of the forecast errors: as an example, Jung and Leutbecher (2008) apply a scale-separation verification approach and quantify the contribution of planetary, synoptic and sub-synoptic scales to the total skill. They analyse the scale dependency of the spread-skill relationship and find that the ECMWF ensemble is over-dispersive at the synoptic scales, with maximum spread and error corresponding to the North Atlantic and North Pacific storm track. More recently, Buizza and Leutbecher (2015) investigated the effect of spatial and temporal filtering on forecast skill, showing that while instantaneous, grid-point fields have forecast skill out to between 16 and 23 days, large-scale, low frequency filtered fields have skill even beyond this range. Note that these conditional and spatial verification approaches, despite being discussed in this generic section on summary performance measures, are process-informative and can obviously also be used for model diagnostics.

Detection of flow-dependent errors (e.g., with Hovmoller diagrams) and spatial verification approaches (alongside traditional verification approaches) can be crucial also for assessing the impacts of improved polar prediction on the predictability of mid-latitude weather. Given the response-time of coupled ocean-atmospheric numerical systems, long lead-time (ensemble) forecasts (10 to 30 days) are needed to explore the linkage between polar regions and mid-latitudes.

All verification scores / model comparisons should be accompanied with confidence intervals and/or significance tests. When comparing different models (or different configurations / schemes /

parameterizations of the same model) it is preferable to perform the inference on the difference of the verification scores for paired samples. Inference on verification results can be performed either by traditional parametric tests (Wilks, 2011, chapter 5; von Storch and Zwiers 1999, chapter 6; Jolliffe, 2007) or by re-sampling and permutation tests and bootstrapping (Efron and Tibshirani, 1993; Gilleland, 2010). Non-parametric re-sampling methods (e.g. bootstrapping) provide an intuitive and distribution-free approach for performing statistical inference on verification results. Figures 2, 4, 5 show some examples of significance tests for the verification of the CMC/ECCC Global Deterministic Prediction System (GDPS).

3.3 Physically-meaningful user-oriented verification for sea-ice prediction

Sea ice models play a key role in environmental prediction for polar regions, by providing ice products for polar marine users as well as a boundary forcing factor for atmospheric prediction. Sea-ice is characterized by several attributes and features:

- i) sea-ice concentration (defined as the fractional area covered by sea-ice, e.g. within a model grid-box), and its derivatives;
- ii) sea-ice extent (defined as the total area covered by sea-ice with a sea-ice concentration exceeding a specified threshold) and
- iii) sea-ice edge (defined as the sea-ice extent boundary position);
- iv) sea ice thickness (which plays a central role in predictability as sea-ice operates as a buffer between the ocean-atmosphere interactions);
- v) sea-ice stage of development (which is partially correlated with sea-ice thickness: usually the older the ice, the thicker);
- vi) sea-ice pressure (which is the negative average of the normal ice stresses);
- vii) sea-ice drift (which is mostly determined by the air-ice stress, the ice-ocean stress and the ice interaction term, i.e. the sea-ice rheology)
- viii) sea-ice deformation (which can lead to formation of leads and pressure ridges, or to the opening of polynias);
- ix) sea-ice floes and icebergs.

Many of these attributes (especially sea-ice pressure and icebergs) are critical for navigation safety. In this report, recommendations for sea-ice verification address, on one hand, the needs of the model developers, and, on the other hand, they target the maritime transport sector (safety of high latitude navigation), from an end-user perspective.

Sea-ice is characterized by a coherent spatial structure, with sharp discontinuities and linear features (e.g. leads and ridges), the presence of spatial features (e.g. ice-shelves and icebergs), and a multi-scale structure (e.g. agglomerates of floes of different sizes). Several satellite-based products are available (e.g. for sea-ice concentration and thickness) and can provide spatially-defined sea-ice observations. Sea-ice can benefit from the enhanced diagnostic power of spatial verification approaches. In Appendix A we provide a concise review and general framework for existing spatial verification techniques. In what follows we suggest some specific spatial verification methods (in addition to traditional verification approaches) for each of the above-mentioned sea-ice attributes and features.

Sea ice concentration is the sea-ice covered areal fraction: it is a continuous variable which ranges in the interval $[0,1]$, where a value of zero corresponds to open water, and a value of one corresponds to a

sea that is fully ice covered. Sea-ice concentration is characterized by a U-shaped distribution, which becomes a uniform distribution as we exclude its extremes (i.e., open water and full ice). As a baseline, we recommend verification of sea-ice concentration using traditional continuous and categorical verification scores (e.g. Lemieux et al, 2016). Seasonal forecasts and climate projections focus mainly on the extent of the whole sea-ice pack. Forecasts at shorter lead times (e.g. 48 hours), on the other hand, are more interested in capturing the sea-ice evolution within the Marginal Ice Zone (MIZ), which is the transition region between ocean open-water and full sea-ice cover. The MIZ is the region where the “action” takes place, including sea-ice freeze-ups and melt-downs, and it corresponds to the sea-ice concentration values of the U-shaped distribution belonging to (0,1), excluding its extremes. Traditional verification statistics evaluated over the whole sea-ice concentration values are bound to be dominated by the extremes of the U-shaped distribution (i.e. open water and full sea-ice coverage). For short-range forecasts, in order to obtain more meaningful statistics, verification of sea-ice concentration should focus on the MIZ, and exclude the majority of open-water and full sea-ice covered grid-points. In order to restrict verification to the MIZ, as an example, scores can be evaluated solely for grid-boxes where (gridded) observations and model have changed with respect to the previous day or week (e.g. van Woert et al, 2004). Following this approach, it is natural (and recommended) to compare verification results against persistence.

Verification of sea-ice concentration with categorical scores requires thresholding. An issue associated with thresholding is that the natural threshold used to distinguish between ice and water can be different in gridded observation products with respect to the model. As an example, Smith et al (2016) verify sea-ice concentration from the Canadian Global Ice Ocean Prediction System (GIOPS) model versus the ice-extent produced by the NOAA IMS (Helfrich et al, 2007), and show the sensitivity of categorical verification scores to the threshold choice. The IMS ice-extent is defined by using a 40% threshold of the NOAA satellite-based sea-ice concentration gridded analysis. The natural thresholding for GIOPS, on the other hand, is 20% sea-ice concentration, because within GIOPS a sea-ice concentration smaller than 20% is associated (during the ocean assimilation process) with above freezing sea surface temperatures, whereas a concentration greater than 20% is associated with below freezing sea surface temperatures. The use of multi-categorical verification scores, where multiple thresholds are considered, can help address (at least partially) this issue. A multi-category contingency table can be evaluated based on different user-relevant thresholds. The entries of this table are then combined and weighted by the entries of a scoring matrix which is defined to balance-out rewards and penalties, while accommodating different users’ perspectives.

Thresholding of sea-ice concentration leads to the definition of sea-ice extent and of the sea-ice edge. Categorical approaches are the natural verification method to analyse these sea-ice attributes (in fact, categorical verification of sea-ice concentration is, *de facto*, verification of the sea-ice extent). Issues associated with the thresholding (e.g., sensitivity of the verification results to the threshold choice) also affects the verification of sea-ice extent and sea-ice edge.

Sea-ice extent and sea-ice edge are more naturally verified spatially. Distance measures for binary images, such as the mean distance and metrics from the Hausdorff family, have been used to verify sea-ice extent (Dukhovskoy et al., 2015). Distance measures from pattern recognition and edge detection theory, such as the Fréchet distance (Heinrichs et al., 2006) or simply geographical distance measures (Hebert et al., 2015) have been used to verify the ice-edge location (Figure 6). These metrics provide

physically meaningful and easy-to-interpret verification results (i.e., a distance in km), and therefore they are particularly suitable for user-relevant applications.

Most of the current operational sea-ice prediction systems are designed to represent the evolution of the sea-ice concentration as a whole, rather than explicitly resolving individual floes, and solely a few sea-ice models start representing the evolution of floe-size distribution (e.g., Horvat and Tziperman, 2015). As Lagrangian particle-based sea-ice models develop towards explicitly resolving size and evolution of individual floes (e.g., Rabatel et al, 2015), spatial verification techniques which enable assessment of the complex multi-scale structure of floes (as an example, scale-separation methods, or feature-based approaches as MODE or SAL) could also be explored for the verification of sea-ice concentration; given the sharp discontinuities and presence of (often uncountable small-scale) features, neighbourhood methods (such as the Fraction Skill Score) could be used to avoid double penalties while accounting for the small drift errors; finally, field-deformation approaches (e.g., DAS or image warping) could be used to quantify these small sea-ice drift errors. Note that with the neighbourhood methods, deterministic sea-ice forecasts can become probabilistic products (e.g., Theis et al, 2005), and probabilistic scores can be used for their evaluation.

Goessling et al (2016) introduce the integrated ice-edge error (IIEE), a user-relevant and intuitive verification measure for assessing sea-ice edge and extent. The IIEE is defined as the area where the forecast and the observations disagree on the ice concentration being above or below 15%. The IIEE can be decomposed into an absolute extent error (AEE, corresponding to the common sea ice extent error), and a misplacement error ($ME = IIEE - AEE$). This approach bridges the traditional categorical scores to spatial verification approaches which quantify displacement errors (such as the object-oriented methods) and provides simple yet informative verification results. Moreover, a probabilistic metric for the verification of contours, termed the spatial probability score (SPS), has been introduced recently as the spatial integral of local Brier Scores (Goessling, personal communication). When applied to deterministic ice edge forecasts, the SPS is reduced to the IIEE, allowing to verify deterministic and probabilistic sea ice forecasts in a common framework.

Sea-ice thickness is continuous and positive-defined (its values are bounded at the lower end by zero), and is characterized by a mixed (zero versus non-zero) and right skewed distribution. Spatially, sea-ice thickness exhibits a coherent spatial structure which can be smooth (e.g. in correspondence with land fast ice), but can also exhibit spatial discontinuities (e.g. in regions of strong convergence and hence of significant deformation, such as north of Canada and north of Greenland) and can be affected by the presence of linear kinematic features (e.g. ridges). Observations of sea-ice thickness include in-situ measurements but also thickness estimates from satellite-based sensors: the latter offer the potential of applying spatial verification approaches.

Verification of sea-ice thickness presents similar characteristics and challenges as verification of precipitation fields. As an example, traditional point-by-point continuous verification scores can be dominated by the few large thickness values (this is especially true for statistics defined by a quadratic rule, such as the MSE) and can be heavily affected by double penalties associated with small sea-ice drift errors. Categorical verification of sea-ice thickness mitigates the effect of the large values. Spatial verification approaches, such as neighbourhood methods, can help address the double penalty issue, and field deformation approaches can help quantify the contribution of the small drift errors. Numerical

models can predict the sea-ice thickness distribution within a grid-cell: where observations support this (e.g., at a super-sites), sub-grid sea-ice thickness distributions can be verified by comparing the moments of these (observed and predicted) distributions, or by using probabilistic and ensemble verification scores, such as the Continuous Ranked Probability Score, or by using statistics which measure the distance between two sample distributions, such as the Kolmogorov-Smirnov distance.

Sea-ice pressure represents perhaps the most critical variable for navigation safety. Sea-ice pressure is predicted by sea-ice prediction systems as a continuous variable, however ship observations report sea-ice pressure in categories, such as beset, severe, moderate, light, and absent. Multi-categorical verification scores are the natural approach for sea-ice pressure verification, where the model-produced continuous values are first calibrated and then thresholded, to be compared to the observed categories. Verification of sea-ice pressure presents several challenges, mainly associated with the observation procedures. As an example, the aforementioned categories are subjective and dependent on the type of ship reporting (e.g., a small vessel versus a large ice-breaker) and the number of categories in the reporting procedures can vary: these aspects add complexity in the model calibration procedure. Moreover, ships often do not report if no pressure was encountered: this practice introduces a sampling bias in the contingency table entries (the categories associated with no observed pressure are under-represented with respect to reality) and can invalidate the verification results. Quantitative (continuous) observation of sea-ice pressure are provided by sea-ice buoys: verification against these can help overcome some of the aforementioned issues, however the currently deployed buoys are very limited in number to achieve a representative spatial coverage and significant verification results. Finally, verification results of sea-ice pressure can suffer from severe representativeness issues. In fact, sea-ice pressure is highly discontinuous in space and can vary horizontally at the meter scale: the ice pressure exerted on a ship haul or measured by a in-situ stress sensor is localized and represents a subgrid scale phenomena, when compared to the model-simulated pressure at the scale of a grid cell (on the order of a few km). Evaluation of sea-ice pressure would benefit from the development of downscaling methods.

Sea-ice stage of development is usually expressed in categories such as nilas and new ice, grey ice and grey-white ice (young ice), (thin, medium, thick) first-year ice, second-year and multi-year ice. Multi-categorical scores are the natural verification approach for sea-ice stage of development. In satellite-based products the sea-ice stage of development is estimated from sea-ice optical properties: the assumptions behind the retrieval algorithms and sea-ice age classification introduce non-negligible uncertainties, and their effects on the verification results should be quantified. Spatially, sea-ice stage of development has similar characteristics as sea-ice concentration, and hence similar spatial verification approaches could be considered.

Sea-ice drift is a vector field, hence each grid-point value is characterized by a speed and a direction. Speed values are positive and right skewed, whereas directions are uniformly distributed in the range $[0^\circ, 360^\circ]$. Spatially, sea-ice drift has a fairly smooth coherent spatial structure, but can exhibit sharp discontinuities (e.g. caused by strong winds or currents, along shear lines, or along the offshore edge of landfast ice). Sea-ice observed drift trajectories are either measured by drifting buoys, or can be derived by using satellite imageries. Buehner et al (1997) use a variational approach to estimate the displacement field between two satellite images of sea-ice. Komarov and Barber (2014) introduce an automated sea-ice tracking system which detect sea-ice drift fields from consecutive RADARSAT

Synthetic Aperture Radar images (Figure 7). Traditional ice-drift verification is performed comparing point-wise speed and direction values: however this approach does not account for the intrinsic correlation existing between nearby grid-point values, and can lead to double penalty errors. Neighbourhood verification approaches (e.g. Marsigli et al, 2005) can avoid the double penalty issue, and field-deformation techniques (e.g., DAS or image warping) could provide more informative feedback on the drift error: they both should be considered for the verification of sea-ice drift trajectories.

In summary, given the challenging spatial characteristics of sea-ice, and treasuring from the available satellite-based spatially-defined sea-ice observation products, we strongly recommend the use of spatial verification approaches (along with traditional verification approaches) for the verification of sea-ice predictions. In particular, the following classes of spatial verification approaches have been identified:

- Distance measures for binary images, pattern recognition and edge detection (such as the Hausdorff, Baddeley and Fréchet distances) are suitable to verify sea-ice extent and/or assess the distance between the forecast and observation ice-edge. These metrics provide physically meaningful and easy-to-interpret verification results (e.g., a distance in km), and therefore they are recommended for user-relevant verification.
- Field deformation (morphing) techniques are already used in sea-ice assimilation and tracking to analyse the displacement of the sea-ice field. Some of the field-deformation techniques (e.g., DAS or image warping) could be exploited for the verification of sea-ice concentration, thickness and sea-ice drift trajectories.
- Feature-based verification techniques (e.g. CRA, MODE or the SAL score) are suggested for verifying the motion and displacement of isolated icebergs and large floes.
- As (Lagrangian particle-based) sea-ice models develop towards explicitly resolving size and evolution of individual floes (as opposed to floe-size distribution), feature-based techniques (e.g. SAL and MODE), along with scale-separation techniques, can also be exploited for assessing the ability of the numerical prediction systems in reproducing the sea-ice floe multi-scale structure.
- Neighborhood verification approaches (e.g. the Fraction Skill Score) can be applied to the verification of sea-ice concentration, thickness and drift trajectories, to avoid double penalties while accounting for small sea-ice drift errors. Neighbourhood approaches enable the treatment of deterministic sea-ice forecasts as probabilistic products, so that probabilistic scores can be used for their evaluation.

A review of spatial verification techniques (with associated key references) is given in Appendix A.

Sea-ice is one of the most obvious indicator of climate change (Stroeve et al., 2007; 2012). Satellite observations of the Arctic sea ice cover are systematically produced since the late 1970s. These observations show that the Arctic minimum sea ice extent (which occurs in mid-September) exhibits a significant downward trend since then. The sea-ice depletion is clearly visible also in seasonal outlooks, further than in decadal and climate projections. The presence of secular trends in verifying data can affect verification results. As an example, Sigmond et al (2013) show that the skill (as measured by the correlation coefficient) of the seasonal forecast of Arctic sea-ice is dominated by the

downward trend of sea-ice area associated with global warming. This result occurs because the correlation coefficient is a measure of linear dependence, and when the observed and forecast variables are dominated by a similar (increasing or decreasing) trend, the correlation coefficient will be also dominated by such a trend, and will exhibit high skill. Sigmond et al (2013) show that once the trend is removed from the forecast and verifying observation, the skill in predicting the variation of the seasonal sea-ice area is significantly reduced. Sea-ice verification practices must account for not-negligible climate trends, especially if the focus of the verification is the assessment of season-by-season accuracy (associated with the inter-annual variability), beyond the secular trend.

4. Conclusions

Verification in polar regions is a complex problem, presenting many challenges and, therefore, many opportunities for research and new development. In this report we have identified some possible new avenues and key research foci in verification methodology, including:

1. Development of verification approaches which account for observation uncertainties (due to measurements, retrievals and gridding algorithms, spatio-temporal sparseness and sampling uncertainty). Observation uncertainty is a non-negligible component of the forecast error, especially in polar regions.
2. Assessment of the impact of using model-dependent analyses (as opposed to observations) for verification; this includes studies which further explore the feasibility of using a multi-analysis approach for verification purposes, and considering representativeness and observation uncertainty estimates from data assimilation.
3. Design of specific process-based model diagnostics to identify shortcomings in the numerical model representation of key polar processes. Model diagnostics usually aim to analyse all physical aspects in representing a few well observed case studies, and involve verification of process-specific physical variables (e.g. radiation, momentum and energy fluxes): these physical processes and variables need to be identified, and the verification strategy needs to be outlined, in close collaboration with model developers.
4. Exploration of the use of multi-variate statistics, conditional (e.g., process-specific) verification and spatial verification (e.g., scale-separation and field-deformation approaches) to better understand the physical nature and sources of model systematic errors in the polar regions.

Sea-ice represents a key variable in polar regions both for numerical modeling (as it provides a boundary forcing for both atmospheric and ocean prediction, and acts as a buffer regulating ocean-atmosphere interactions), and for decision-makers and end-users (e.g. to guarantee safety in the maritime transport sector). Sea-ice is also one of the variables most strongly affected by climate change. Sea-ice is characterized by a range of attributes (e.g. concentration, extent, thickness, pressure, location of the edge, drift), which are each associated with specific verification challenges. Sea-ice verification can benefit from the existence of several satellite-based spatially-defined sea-ice observation products, which enable the application of spatial verification approaches. Spatial verification approaches address some of the issues associated with traditional point-by-point verification (e.g., double penalties) and have enhanced diagnostic power (e.g., they can assess distance errors in km): spatial verification approaches can provide user-relevant, informative and meaningful verification diagnostics for sea-ice prediction.

Traditional summary verification statistics will be used to monitor and compare different numerical models contributing to YOPP, and to characterize predictability limits in polar regions. For the basic surface and upper-air atmospheric variables, YOPP should meet the CBS verification standards (Tables 1a,b). Where YOPP benefits from more frequent surface and/or upper-air observations, verification initial times and frequencies should be adapted accordingly. Traditional continuous and categorical scores, and probabilistic verification statistics should be complemented with extreme dependence indices and the more recent ensemble verification statistics. Simple, yet informative, summary statistics and meaningful graphical displays are essential for a correct interpretation of the verification results. Aggregation of the verification results over a large data sample helps attaining statistically robust verification statistics. However this must be complemented by the application of physically meaningful stratifications. Inference (statistical significance and confidence intervals) of verification results should be regarded as a standard procedure. Non-parametric re-sampling methods (e.g. bootstrapping with suitable choice of block length) provide an intuitive and distribution-free approach for performing statistical inference on verification results.

Apart from sea-ice, this report has not focused on specific variables but attempted to characterize suitable verification methodologies in more general terms. However, one critical variable which will require more study (with tailored verification approaches) in the forthcoming years is snow at the surface: similarly to sea-ice, this variable has crucial importance both for modeling (as it regulates all the land-atmosphere flux exchanges, and the albedo and radiation budget) and from an end-user perspective (e.g. for land transport). Representation of clouds (especially low-level mixed-phase clouds) in the polar regions is also a key process which needs to be specifically addressed (the physical feedback on the radiation budget due to the presence of clouds is possibly the most uncertain feedback process in climate modeling): recent advancements in satellite observations of cloud composition offer the potential of enhanced verification diagnostics for cloud representation in numerical prediction systems. Increased attention in forthcoming years should be also given to the validation of how numerical models reproduce teleconnections, stratosphere to troposphere propagation, and linkages between polar and mid-latitude weather. A simple example is to use Hovmöller diagrams to detect flow-dependent error propagation; similarly, backtracking can identify the sources of such flow-dependent error. More sophisticated verification efforts could explore a conditional verification based on specific weather regimes (e.g. by applying cluster analysis to self-organizing maps and/or analysing composites), analyse how weather features in different geographical regions and at different times are intercorrelated, and verify whether numerical models correctly represent these inter-correlations.

For further reading on forecast verification we suggest Jolliffe and Stephenson (2012) and the verification chapter in Wilks (2011) for basic concepts and traditional approaches. The verification webpage <http://www.cawcr.gov.au/projects/verification> maintained by the WMO Joint Working Group on Forecast Verification Research (JWGFVR) includes a concise review of basic and more advanced techniques, and provides a vast list of references and links to verification-related web-sites. The WMO web-page https://www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html includes past JWGFVR publications with recommendations for the verification of clouds and precipitation forecasts. The MesoVICT web-page <http://www.ral.ucar.edu/projects/icp/references.html> includes a list of more than 200 peer-reviewed publications on spatial verification techniques. The Meteorological Evaluation Tools (MET, <http://www.dtcenter.org/met/users/>) is a freely available verification package developed at NCAR which includes traditional verification methods (continuous, categorical, probabilistic and

ensemble) and spatial verification techniques, supports GRIB and netcdf files, handles operational size data loads, and could be used for YOPP verification purposes. Finally, the review articles of Ebert et al (2013) and Ebert and Brown (2015) outline the most recent advancements and future challenges in verification research.

Acknowledgements:

B. Casati wish to thank A. Zadra, P. Vaillancourt, P. Pestieau, T. Carriers, G. Smith, M. Buehner, P. Houtekamer, E. Ebert, H. Goessling and P. Bauer for the scientific discussions and revisions, which significantly improved this document.

References

- Ahrens, B. and S. Jaun, 2007: On evaluation of ensemble precipitation forecasts with observation-based ensembles. *Adv. Geosci.*, **10**, 139-144.
- Alexander, G. D., J. A. Weinman, and J. L. Schols, 1998: The use of digital image warping of microwave integrated water vapor imagery to improve forecasts of marine extratropical cyclones. *Mon. Wea. Rev.*, **126**, 1469-1496.
- Alt H. and Godau M., 1995; Computing the Fréchet distance between two polygonal curves, *Internat. J. Comput. Geom. Appl.*, **5**:75-91.
- Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.*, **8**, 401-417.
- Baddeley, A. J., 1992a: Errors in binary images and an L_p version of the Hausdorff metric. *Nieuw Arch. Wiskunde*, **10**, 157-183.
- Baddeley, A. J., 1992b: An error metric for binary images. *Robust Computer Vision: Quality of Vision Algorithms*, W. Förstner and S. Ruwiedel, Ed., Wichmann, 59-78.
- Bauer, P., L. Magnusson, J.-N. Thepaut, and T. M. Hamill, 2014: Aspects of ECMWF model performance in polar areas. *Q. J. R. Meteorol. Soc.*, **142**, 583-596.
- Bowler, N.E., 2008: Accounting for the effect of observation errors on verification of MOGREPS. *Meteorol. Appl.*, **15**, 199-205.
- Briggs, W.M. and R.A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.*, **125**, 1329-1341.
- Buehner, M, K.R. Thompson, I. Peterson, 1997: An Inverse Method for Tracking Ice Motion in the Marginal Ice Zone Using Sequential Satellite Images. *J. of Atmos and Oceanic Technology*, **14**, 1455-1466.
- Buizza, R. and M. Leutbecher, 2015: The forecast skill horizon. *Q. J. R. Meteorol. Soc.*, **141**, 3366–3382.
- Candille, G., C. Côté, P. L. Houtekamer and G. Pellerin, 2007: Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.*, **135**, 1140-1147.
- Casati, B., Ross, D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts, *Meteorol. Appl.*, **11**, 141-154.
- Casati, B., and L.J. Wilson, 2007: A New spatial-scale decomposition of the Brier score: Application to

- the verification of lightning probability forecasts. *Mon. Wea. Rev.*, **135**, 3052-3069.
- Casati, B., 2010: New developments of the intensity-scale technique within the Spatial Verification Methods Intercomparison Project. *Wea. Forecasting*, **25**, 113-143.
- Casati, B., V. Fortin, and L. Wilson, 2014: A wavelet-based verification approach to account for the variation in sparseness of gauge observation networks. *World Weather Open Science Conference*, Montreal, Canada, 16-21 August 2014.
- Christensen, H. M., I. M. Moroz, and T. N. Palmer, 2014: Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts. *Q. J. R. Meteorol. Soc.*, **141**, 538–549.
- Ciach G.J., and W.F. Krajewski, 1999. On the estimation of radar rainfall error variance. *Adv. Water Resources*, **22**, 585-595.
- Crocker, R, R. Neal, D. Fereday, 2014: Verifying deterministic and probabilistic forecasts of objectively clustered weather regimes. *World Weather Open Science Conference*, Montreal, Canada, 16-21 August 2014.
- Davis, C. A., B. G. Brown, and R. G. Bullock, 2006: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772 - 1784.
- Davis, C. A., B. G. Brown, and R. G. Bullock, 2006: Object-based verification of precipitation forecasts, Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785 - 1795.
- Davis, C.A., B.G. Brown, R.G. Bullock and J. Halley Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Wea. Forecasting*, **24** (5), 1252 - 1267.
- Day, J.J., S. Tietsche, E. Hawkins, 2014: Pan-Arctic and regional sea ice predictability: initialization month dependence. *Journal of Climate*, **27**, 4371-4390.
- de Elia, R., R. Laprise, and B. Denis, 2002: Forecasting skill limits of nested, limited-area models: A perfect-model approach. *Mon. Wea. Rev.*, **130**, 2006-2023.
- Denis, B., R. Laprise, D. Caya, and J. Côté, 2002: Downscaling ability of one-way nested regional climate models: the Big-Brother Experiment. *Climate Dynamics*, **18**, 627-646.
- Denis, B., R. Laprise, D. Caya, 2003: Sensitivity of a Regional climate Model to the resolution of the lateral boundary conditions. *Climate Dynamics*, **20**, 107-126.
- De Sales, F. and Y. Xue, 2010: Assessing the dynamic-downscaling ability over South America using the intensity-scale verification technique. *Int. J. Climatol.*, **31** (8), 1205-1221.
- Desroziers G, Berre L, Chapnik B, Poli P. 2005. Diagnosis of observation, background, and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.* **131**, 3385-3396.
- Dubuisson, M-P. and A.K. Jain (1994): A Modified Hausdorff Distance for Object Matching. *Proc. International Conference on Pattern Recognition*, Jerusalem, Israel, pp 566-568.
- Dukhovskoy, D. S., J. Ubnoske, E. Blanchard-Wrigglesworth, H. R. Hiester, and A. Proshutinsky, 2015: Skill metrics for evaluation and comparison of sea ice models. *J. Geophys. Res. Oceans*, **120**, 5910-5931.
- Ebert, E. E. and J. L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.

- Ebert (2008) "Fuzzy verification of high resolution gridded forecasts: a review and proposed framework", *Meteorol. Appl.*, **15**, 51-64.
- Ebert, E.E. and W.A. Gallus, Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting.*, **24** (5), 1401-1415
- Ebert, E., L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. Göber, S. Joslyn, B. Brown, T. Fowler, and A. Watkins, 2013: Progress and challenges in forecast verification. *Meteorol. Appl.*, **20**: 130-139.
- Ebert, E. and B. Brown, 2015: Numerical Prediction of the Earth System: cross-cutting research on verification techniques. *Seamless Prediction of the Earth System: from minutes to Months*, WMO-no. 1156.
- Efron, B. and R.J. Tibshirani, 1993: *An introduction to the bootstrap*. Chapman and Hall, 436 pp.
- Eiter, T., and H. Mannila, 1994: Computing Discrete Fréchet Distance. *Technical report of Christian Doppler Laboratory*; Vienna University of technology, num. CD-TR 94/64.
- Ferro, C.A.T., and D.B. Stephenson, 2011: Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699-713.
- Germann, U., and I. Zawadzki, 2004: Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteorol.*, **43**, 74 - 89.
- Gilleland, E., 2010: Confidence intervals for forecast verification. *NCAR Technical Note*, NCAR/TN-479+STR, 71pp.
- Gilleland, E., 2011: Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases. *Wea. Forecasting*, **26** (3), 409-415.
- Gilleland, E., D. Ahijevych, B.G. Brown, and E.E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365-1373.
- Gilleland, E., J. Lindström, and F. Lindgren, 2010: Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Wea. Forecasting*, **25** (4), 1249 - 1262.
- Gilleland, E., T. C. M. Lee, J. Halley Gotway, R. G. Bullock, and B. G. Brown, 2008: Computationally efficient spatial forecast verification using Baddeley's Δ image metric. *Mon. Wea. Rev.*, **136** (5), 1747-1757.
- Goessling, H. F., S. Tietsche, J. J. Day, E. Hawkins, and T. Jung, 2016: Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.*, **43**, 1642-1650.
- Grams, J. S., W. A. Gallus Jr., S. E. Koch, L. S. Wharton, A. Lough, and E. E. Ebert, 2006: The Use of a Modified Ebert-McBride Technique to Evaluate Mesoscale Model QPF as a Function of Convective System Morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288-306.
- Haiden T., H. Kabelwa, M. Mittermaier, A. Okagaki, T. Robinson, 2014: Draft document of standardized surface verification of deterministic NWP products. To be included in 2016 in the new CBS manual WMO-No. 485.
- Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier, J. J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeorology*, **2**, 406-418.
- Hebert D.A., R. A. Allard, E.J. Metzger, P.G. Posey, R.H. Preller, A.J. Wallcraft, M.W. Phelps, and O. M. Smedstad, 2015: Short-term sea ice forecasting: An assessment of ice concentration and ice drift forecasts using the U.S. Navy's Arctic Cap Nowcast/Forecast System. *J. Geophys. Res.*

Oceans, **120**, 8327-8345.

- Heinrichs, J.F., D.J. Cavalieri, and T. Markus 2006: Assessment of the AMSR-E Sea Ice Concentration Product at the Ice Edge Using RADARSAT-1 and MODIS Imagery. *IEEE transactions on Geoscience and Remote Sensing*, **44**(11), 3070-3080.
- Helfrich, S.R., D. McNamara, B.H. Ramsay, T. Baldwin, T. Kasheta, 2007. Enhancements to, and forthcoming developments in the interactive Multisensor Snow and Ice Mapping System (IMS). *Hydrol. Process.*, **21**, 1576-1586.
- Hoffman, R. N. and C. Grassotti, 1996: A technique for assimilating SSM/I observations of marine atmospheric storms: Tests with ECMWF analyses. *J. Appl. Meteorol.*, **35**, 1177-1188.
- Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758 - 2770.
- Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I. The wind field. *Tellus* **38A**: 111-136.
- Horvat, C. and R. Tziperman, 2015: A prognostic model of the sea-ice floe size and thickness distribution. *The Cryosphere*, **9**, 2119-2134.
- Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637-650.
- Jolliffe, I.T., and D.B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition*. Wiley and Sons Ltd, 274 pp.
- Jung, T. and Leutbecher, 2007: Performance of the ECMWF forecasting system in the Arctic during winter. *Q. J. R. Meteorol. Soc.*, **133**, 1327-1340.
- Jung, T. and Leutbecher, 2008: Scale dependent verification of ensemble forecasts. *Q. J. R. Meteorol. Soc.*, **134**, 973-984.
- Jung, T. and M. Matsueda, 2014: Verification of global weather forecasting systems in polar regions using TIGGE data. *Q. J. R. Meteorol. Soc.*, **142**, 574-582.
- Keil, C. and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.*, **135**, 3248-3259.
- Keil, C. and G.C. Craig, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting*, **24** (5), 1297-1308.
- Komarov, A.S., and D.G. Barber, 2014: Sea ice motion tracking from sequential dual-polarization RADARSAT-2 images. *IEEE Transactions on Geoscience and Remote Sensing*, **52**(1), 121-136.
- Lack, S., G.L. Limpert and N.I. Fox, 2010: An object-oriented multiscale verification scheme. *Wea. Forecasting*, **25**, 79-92.
- Lemieux, J.F. and co-authors, 2016: The Regional Ice Prediction system (RIPS): verification of forecast sea ice concentration. *Q.J. Royal Meteorol. Soc.*, **142**, 632-643.
- Liu, Z.-Q., and F. Rabier, 2003. The potential of high-density observations for numerical weather prediction. A study with simulated observations. *Q. J. R. Meteorol. Soc.*, **129**: 3013-3035.
- Livina, V. N., N. R. Edwards, S. Goswami, and T. M. Lenton, 2008: A wavelet-coefficient score for comparison of two-dimensional climatic-data fields. *Q. J. R. Meteorol. Soc.*, **134**, 941 - 955.
- Marsigli, C., Boccanera, A. Montani, and T. Paccagnella, 2005: The COSMO-LEPS ensemble system: validation of the methodology and verification. *Nonlinear Proc. Geoph.*, **12**, 527 - 536.

- Marzban, C., and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Wea. Forecasting*, **21** (5), 824 - 838.
- Marzban, C., and S. Sandgathe, 2008: Cluster Analysis for Object-Oriented Verification of Fields: A Variation. *Mon. Wea. Rev.*, **136** (3), 1013 - 1025.
- Marzban, C., and S. Sandgathe, 2010: Optical flow for verification. *Wea. Forecasting*, **25**, 1479-1494.
- Mittermaier, M.P., 2014: How temporally representative are synoptic observations? *World Weather Open Science Conference*, Montreal, Canada, 16-21 August 2014.
- Nachamkin, J. E., S. Chen, and J. S. Schmidt, 2005: Evaluation of heavy precipitation forecasts using composite-based methods: A distributions-oriented approach. *Mon. Wea. Rev.*, **133**, 2163 - 2177.
- Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.*, **132**, 941 - 955.
- Nehrkorn, T., R. N. Hoffman, C. Grassotti, and J.-F. Louis. 2003: Feature calibration and alignment to represent forecast errors: Empirical regularization. *Q.J.R. Meteorol. Soc.*, **129**, 195-218
- Park, Y-Y, R. Buizza, M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.*, **134**, 2029-2050.
- Peli, T. and Malah, D. (1982) A study on edge detection algorithms. *Computer Graphics and Image Processing*, **20**, 1-21.
- Pratt, W. K., 2001: *Digital Image Processing*. Third edition, John Wiley and Sons, 738 pp.
- Rabatel, M., S. Labbé, and J. Weiss, 2015: Dynamics of an assembly of rigid ice floes. *J. Geophys. Res.*, **120** (9), 5887-5909.
- Roberts, N. M. and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78 - 96.
- Roebber, P.J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601-608.
- Saetra, O., H. Hersbach, J.-R. Bidlot and D.S. Richardson, 2004: Effects of observation errors for ensemble spread and reliability. *Mon Wea. Rev.*, **132**, 1487-1501.
- Santos, C. and A. Ghelli, 2011: Observational probability method to assess ensemble precipitation forecast, *Q.J.R. Meteorol. Soc.*, **138**, 209-221.
- Schwedler, B. R. J. and M.E. Baldwin, 2011: Diagnosing the sensitivity of binary image measures to bias, location, and event frequency within a forecast verification framework. *Wea. Forecasting*, **26**, 1032-1044.
- Sigmond, M., J.C. Fyfe, G.M. Flato, V.V. Kharin and W.J. Merryfield, 2013: Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophys. Res. Lett.*, **40**, 529-534.
- Smith, G.C. and co-authors, 2016: Sea ice forecast verification in the Canadian Global Ice Ocean Prediction System. *Q. J. R. Meteorol. Soc.*, **142**, 659-671.
- Stamus, P.A., F.H. Carr, D.P. Baumhefner, 1992: Application of a scale-separation verification technique to regional forecast models. *Mon Wea. Rev.*, **120**, 149-163.
- Stephenson D.B., B. Casati, C.A.T. Ferro and C.A. Wilson, 2008: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Appl.*, **15**, 41-50.
- Stroeve, J.C., M.M. Holland, W. Meier, T. Scambos, M.C. Serreze, 2007: Arctic sea-ice decline: faster than forecast. *Geophys. Res. Lett.*, **34**, L09501.

- Stroeve, J.C., M.C. Serreze, M.M. Holland, J.E. Kay, J. Malanik, A.P. Barrett, 2012: The Arctic's rapidly shrinking sea ice cover: a research synthesis. *Climatic Change*, **110**, 1005-1027.
- Talagrand, O., R. Vautard and B. Strauss, 1999: Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability*, Reading, UK.
- Taylor, K.E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106** (D7), 7183-7192.
- Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteorol. Appl.*, **12**, 257 - 268.
- Tustison, B., E. Foufoula-Georgiou, and D. Harris, 2003: Scale-recursive estimation for multisensor quantitative precipitation forecast verification: A preliminary assessment. *J. Geophys. Res.*, **108**, D8, 8377.
- Van Woert, M., C. Zou, W. Meier, P. Hovey, R. Preller, P. Posey, 2004: Forecast Verification of the Polar Ice Prediction System (PIPS) Sea Ice Concentration Fields. *Journal of Atmospheric and Oceanic Technology*, **21** (6), 944-957.
- Venugopal, V., S. Basu, E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.* **110**: D8, D08111
10.1029/2004JD005395.
- von Storch, H. and F.W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.
- Weigel, A.P. and S.J. Mason, 2011: The generalized discrimination Score for ensemble forecasts. *Mon. Wea. Rev.*, **139**, 3069-3074.
- Wernli, H., M. Paulat, M. Hagen and C. Frei, 2008: SAL- a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, **136**, 4470-4487.
- Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Science*. Third edition, Elsevier, 676 pp.
- WMO-No. 485, Manual on the Global Data-Processing and Forecasting System. available at <http://www.wmo.int/pages/prog/www/DPFS/Manual/GDPFS-Manual.html>
- WMO-No. 893, Recommendation 3 (CBS-Ext(98)), Amendments to the Manual on the Global Data-Processing System (WMO-No. 485).
- WMO-No. 1091, Guidelines on Ensemble Prediction Systems and Forecastings. Available at <http://www.wmo.int/pages/prog/www/manuals.html>.
- Yates, E., S. Anquetin, V. Ducrocq, J.-D. Creutin, D. Ricard and K. Chancibault, 2006: Point and areal validation of forecast precipitation fields. *Meteorol. Appl.*, **13**, 1-20.
- Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, **105** (D8), 10,129-10,146.
- Zhu, M., V. Lakshmanan, P. Zhang, Y. Hong, K. Cheng, and S. Chen, 2011: Spatial verification using a true metric. *Atmos. Res.*, **102**, 408-419, doi:10.1016/j.atmosres.2011.09.004.

Appendix A: a concise review of spatial verification methods

Weather variables defined over spatial fields are characterized by a coherent spatial structure and the presence of spatial features. Traditional verification scores evaluated across a grid, with observation and forecast compared by individually matching grid-point by grid-point values, do not account for such spatial structures or the intrinsic spatial correlation existing between nearby grid-points. This practice leads, for example, to double penalties (e.g., associated with small spatial displacements), which become more and more costly as NWP resolution increases (to the extent that high resolution forecasts typically have worse verification scores than their coarser resolution counterparts, despite providing more detailed and realistic features). To address these issues, several new spatial verification approaches have been developed in the past two decades. Although spatial verification methods have initially been developed for precipitation forecasts, these techniques can be applied (and adapted) to many other variables. Spatial verification techniques aim to: i) account for field spatial structure and features; ii) provide information on error in physical terms (meaningful verification); iii) assess location and timing errors (separately from intensity error); and iv) account for small time-space uncertainties (i.e., avoid double-penalty issues).

Spatial verification approaches can be broadly categorized in five classes: scale-separation, neighborhood, field deformation, feature-based, and distance measures for binary images.

1. Scale-separation verification approaches decompose forecast and observation fields into the sum of spatial components on different scales by using a single band spatial filter (e.g., Fourier transforms, wavelets, spherical harmonics); then traditional (continuous, categorical or probabilistic) verification is performed on each spatial scale component, separately. These approaches make it possible to: i) evaluate bias, error and skill on different scales; ii) investigate the scale-dependence of forecast predictability (e.g., determine the no-skill to skill transition scale); iii) assess the forecast versus observation scale structure. Scale-separation techniques provide feedback on physical processes associated with weather phenomena on different scales (e.g. frontal systems versus convective precipitation; planetary, synoptic and sub-synoptic scales). Scale-separation techniques have been applied both to weather forecasts (Briggs and Levine, 1997; Zepeda-Arce et al, 2000; Harris et al, 2001, Tustison et al, 2003; Casati et al, 2004; Casati, 2010; Casati and Wilson 2007; Jung and Leutbecher, 2008) and climate studies (Denis et al 2002,2003; deElia et al 2002; Livina et al, 2008; de Sales and Xue 2010; Stamus et al, 1992).
2. Neighborhood verification approaches (sometimes known as “fuzzy” verification approaches) were specifically developed to reward enhanced resolution in NWP systems. High resolution models provide more realistic forecasts than their coarse-resolution counterparts. Traditional verification approaches based on point-by-point observation-forecast matching, however, do not account for small space-time displacements (leading to a double penalty) and/or observation uncertainties. High resolution forecasts, therefore, tend to score worse than their coarser resolution counterparts, possibly due solely to their higher spatial variability. Neighborhood spatial verification approaches address this issue by relaxing the requirement of exact forecast location (and timing), and define neighborhoods of grid-points (both in space and time), within which the forecast and observation are matched. In this fashion both forecast (small) space-time displacements and observation uncertainties are accounted for. Note that (as opposed to the

single-band spatial filter of scale-separation approaches) neighborhoods are obtained by a low-band spatial filter (i.e., smoothing): as the neighborhood size (scale) increases, the exact obs-forecast matching criteria becomes more and more relaxed. Within the neighborhoods approaches, two types of forecast-observation matching have been identified: if spatial observations are available, then a neighborhood in the forecast field is compared to a neighborhood in the observation field (no-nf); on the other hand, if observations are only available at point-locations (e.g., station observations), then the single observation is matched to a neighborhood of forecast values (so-nf). The data treatments applied within a neighborhood represent different verification strategies: the simplest approach is to up-scale (e.g., average) the variable within the neighborhood, and then perform traditional verification (e.g., Yates et al, 2006); Roberts and Lean (2008) verify the frequency of the event (defined by thresholding precipitation intensities) within the neighborhood by using the Fraction Skill Score; Marsigli et al (2005) verify different aspects of the forecast versus observation marginal distribution (e.g. mean, max, and various quantiles) within the neighborhood; Theis et al (2005) use the distribution of the neighborhood values to define probabilities from the deterministic forecast, and then apply probabilistic verification approaches; Atger (2001) considers forecast precipitation neighborhoods near single stations (so-nf) and evaluates spatial multi-event contingency tables and ROC curves (for different thresholds, distances, timing). Ebert (2008) provides a very interesting review, comparison, and proposed framework for neighborhood verification approaches.

The next two classes of verification techniques (field deformation and feature based) have as common approach the decomposition of the forecast error into displacement and intensity error. The displacement error is obtained by translating the forecast until a best match criterion with the observation is satisfied, whereas the amplitude error is obtained by applying a scalar correction to the translated forecast. Field deformation techniques perform such decomposition over the whole field, whereas feature-based techniques evaluate displacement and intensity error for paired obs-forecast features. These approaches reflect human thinking and eye-ball verification, and are suitable for many end-users because they measure the error in physical terms (e.g. distance errors in km) and provide verification results which are of easy interpretation.

3. Field deformation techniques use a vector field (which can be interpreted as advection or wind field) to deform the forecast field towards the observed field, up until an optimal fit is found (e.g. by maximizing a likelihood function). An amplitude (scalar) field is then applied, in order to correct the intensities of the deformed forecast field to those of the observed field. Field deformation techniques usually perform the error decomposition on different spectral components (as for the scale-separation approaches): hence they directly inform about small scale uncertainty versus large scale errors. Pioneer studies on field-deformation techniques are done by Hoffmann et al (1995); Hoffman and Grassotti (1996), Nehr Korn et al. (2003), and Germann and Zawadzki (2004), with applications in data assimilation and now-casting. Field deformation verification approaches include the Displacement and Amplitude Score (DAS) introduced by Keil and Craig (2007, 2009); optical flow, introduced by Marzbar and Sandgathe (2010); image warping, introduced by Alexander et al (1998) and more developed by Gilleland et al (2010).
4. Feature based verification techniques first identify and isolate (precipitation) features in forecast

and observation fields (e.g., by thresholding, image processing, using composites, cluster analysis, etc.), and then assess the displacement and amount (extent and intensity) of error for each pair of observed and forecast features. Some feature-based approaches also i) evaluate distance-based contingency tables and categorical scores; ii) perform verification as function of feature size (scale); iii) add the time dimension for the assessment of the timing error. Feature-based approaches include the Contiguous Rain Area (CRA) technique, introduced by Ebert and McBride (2000) and further developed by Grams et al (2006) and Ebert and Gallus (2009); the Method for Object-based Diagnostic Evaluation (MODE), introduced by Davis et al (2006a,b; 2009); the Structure Amplitude Location (SAL) score, introduced by Wernli et al (2008); the procrustes approach introduced by Lack et al (2010); the composite method, introduced by Nachamkin (2004, 2005); and the cluster analysis approach, introduced by Marzban and Sandgathe (2006, 2008).

5. Distance measures for binary images were developed in image processing for edge detection and/or pattern recognition, and include the Pratt's (2001) Figure of Merit (FoM), the Fréchet distance (Alt and Godau, 1995; Eiter and Mannila, 1994), the Hausdorff metric and its derivatives, the modified and partial Hausdorff distances (Dubuisson and Jain, 1994), the mean error distance (Peli and Malah, 1982), and the Baddeley Delta metric (Baddeley, 1992a,b). These distance measures are sensitive to the difference in shape and extent of objects. They detect the distance/displacement between forecast and observation features, and are not very sensitive to noise. Because of these characteristics, these metrics seem suitable for the verification of (thresholded) precipitation fields and sea-ice, also when comparing forecasts with different resolutions. Until recently, few studies have exploited these metrics for spatial verification. For example, Schwedler and Baldwin (2011) explore the behaviour of such metrics on a set of idealized cases. Gilleland (2011) analyse the behaviour of the Baddeley Delta metric in the context of the Spatial Forecast Verification Methods Inter-comparison Project (SVx ICP), and Gilleland et al. (2008) use it to merge and match objects within the MODE algorithm. Venugopal et al. (2005) introduce a precipitation verification index that combines continuous verification statistics and the partial Hausdorff distance. Zhu et al. (2011) introduce a binary image metric, defined as the weighted sum of an overlap distance and the mean displacement between the forecast and observation datasets, and apply it to the SVx ICP case studies. In the sea-ice research community, Hebert et al. (2015) use distance metrics to evaluate the sea-ice edge distance; Heinrichs et al. (2006) apply the Fréchet distance for verifying sea-ice edge distance; and Dukhovskoy et al. (2015) explore the suitability of mean error distance, as well as Hausdorff and Modified Hausdorff distances for the verification of the sea-ice extent.

To understand the capabilities (and limits) of these recently developed spatial techniques, the verification research community has coordinated two international spatial meta-verification inter-comparisons: the major outcomes from the first Spatial Verification Inter-comparison Project (SVxICP) are summarized by Gilleland et al. (2010), and in several scientific articles published in a special issue of *Weather and Forecasting* (volumes 24 and 25) and the *Bulletin of the American Meteorological Society* (volume 91). The second (on-going) Mesoscale Verification Inter-comparison in Complex Terrain (MesoVICT, <http://www.ral.ucar.edu/projects/icp>) extends SVxICP by considering several different weather variables (e.g., wind, temperatures, etc) in addition to precipitation, and by including

ensemble predictions in addition to deterministic forecasts. Results of these inter-comparison projects are extremely useful for the research community itself: in part, they reveal the different diagnostic capabilities of the different verification techniques, and often drive the needs for further development. In addition, they provide guidance to end-users in choosing the spatial verification approaches that are most suitable for their needs.

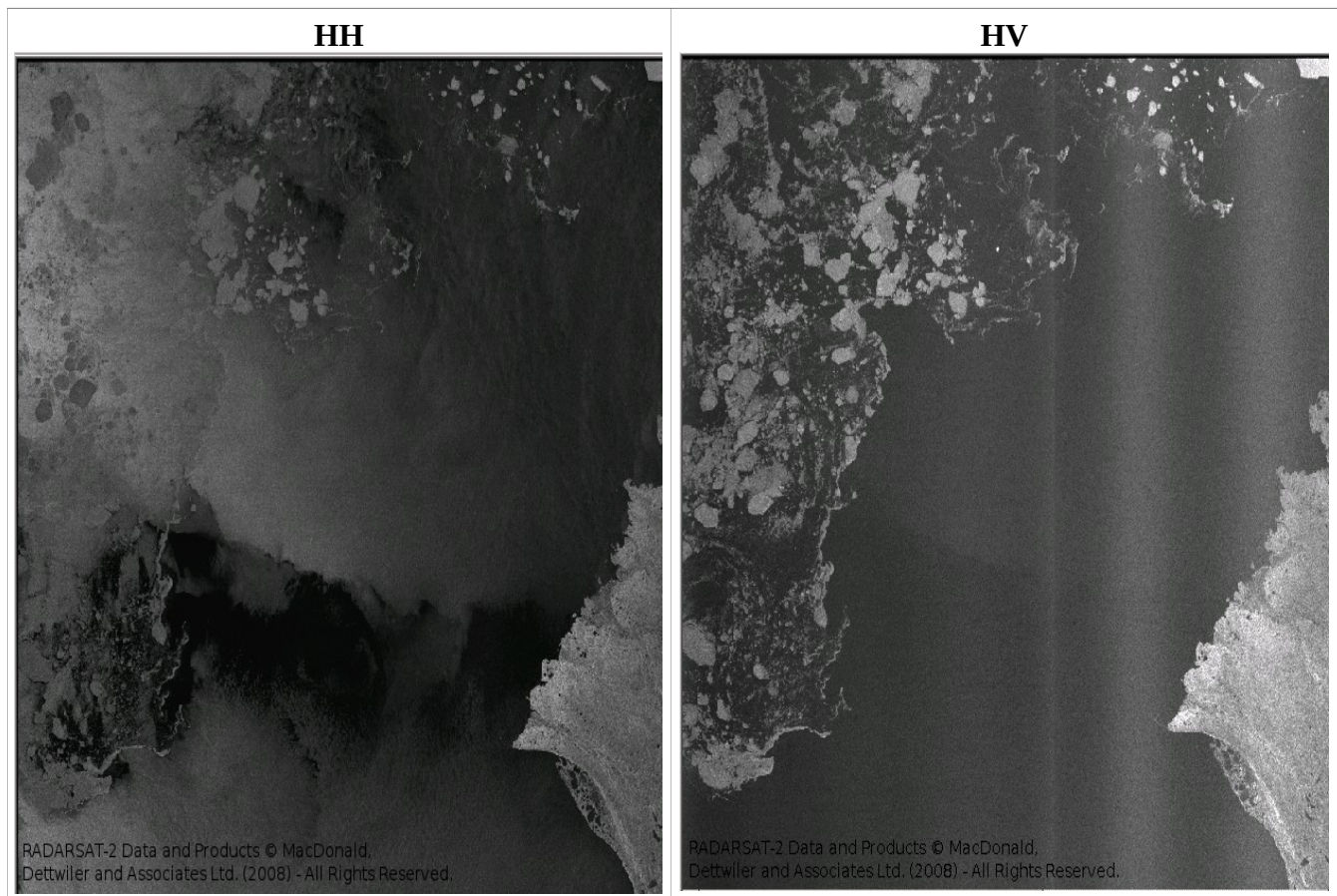


Figure 1: RADARSAT-2 co-polarization (HH, left) and cross-polarization (HV, right) Synthetic Aperture Radar (SAR) images for the retrieval of sea-ice. The SAR technology can see through clouds. Moreover, the cross-polarization shows less incidence-angle dependence than the co-polarization, and is less sensitive to wind effects, so that the HV channel is more reliable than the HH channel for sea ice detection. Images are courtesy of Angela Cheng, Canadian Ice Service / Environment Canada.

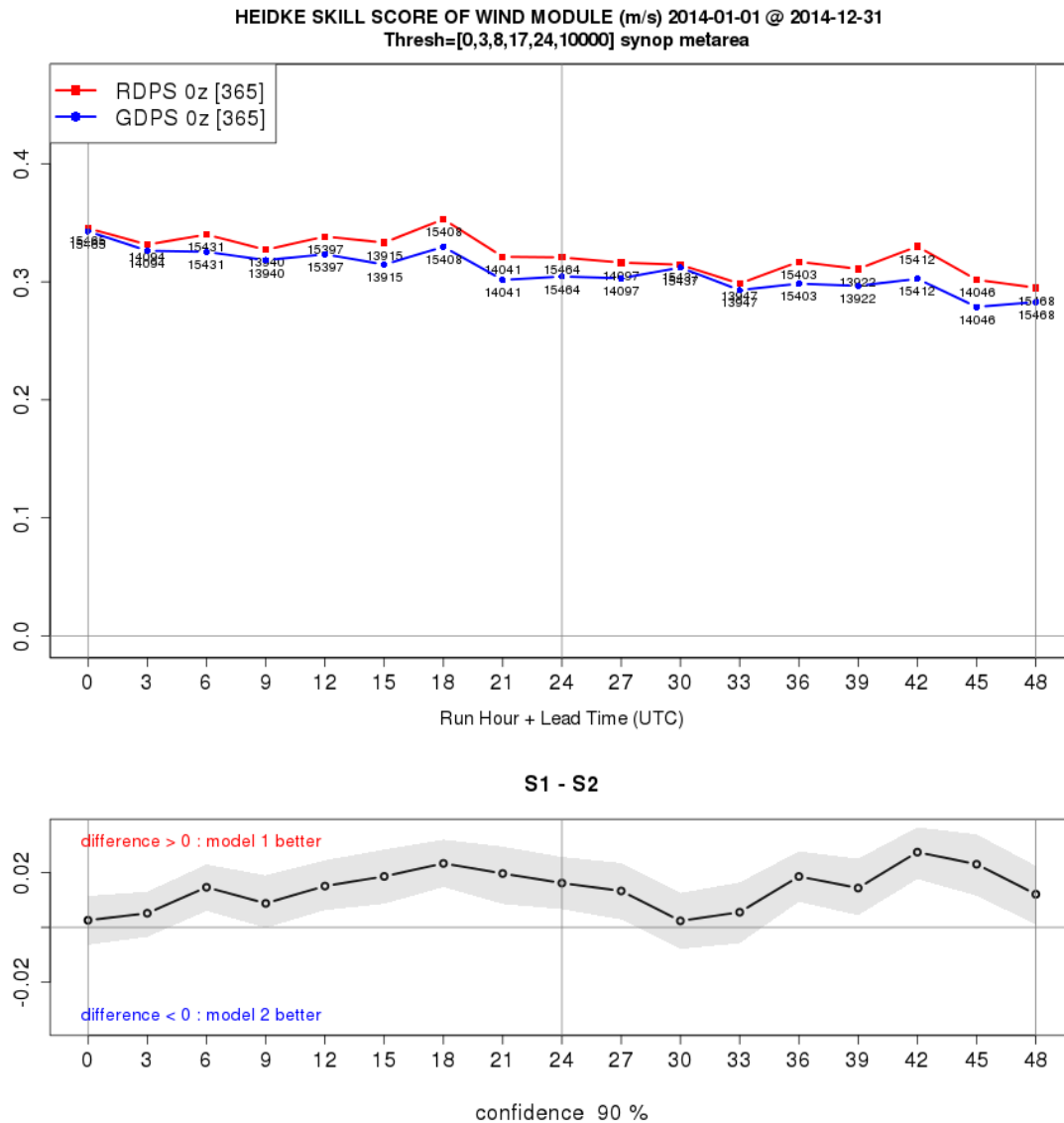


Figure 2: (top) multi-category Heidke Skill Score (HSS) for several wind speed thresholds [0,3,8,17,24,max] for the CMC Regional (red) and Global (blue) Deterministic Prediction Systems as a function of forecast time, evaluated against synoptic observations over the Metarea XVIII domain (covering most of the Canadian Arctic Islands) and (bottom) the corresponding 5-95 % confidence intervals calculated on the difference between the two models, for the period January 1 to December 31, 2014. Confidence intervals are evaluated by block bootstrapping. Image is courtesy of François Lemay and Tom Robinson, Canadian Meteorological Centre / Environment Canada.

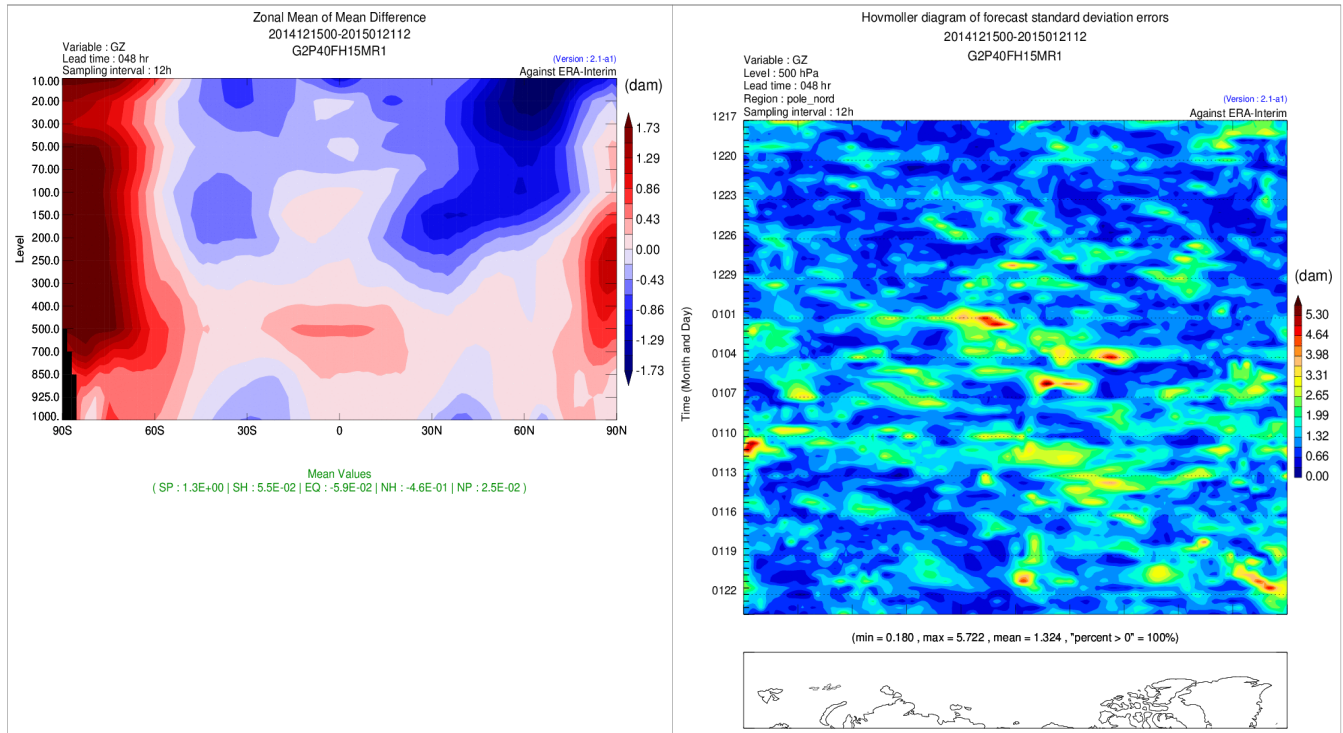


Figure 3: (left) zonally averaged bias of geopotential height at different vertical levels and (right) Hovmoller diagram (longitude-time) of the error standard deviation of geopotential height at 500 hPa meridionally averaged over the North Polar Region, for a 48h lead time experimental run of the CMC global deterministic prediction system versus ERA-interim analysis. Left panel: the geopotential height in the Antarctic is affected by a large positive bias, for all vertical levels; the Arctic is affected by a positive bias in the troposphere; northern mid-latitudes exhibit a strong negative bias in the stratosphere. The Hovmoller diagram displays an example of flow-dependent error propagation. Images are courtesy of Stephane Laroche and C. Charette (RPN/MRD/EC).

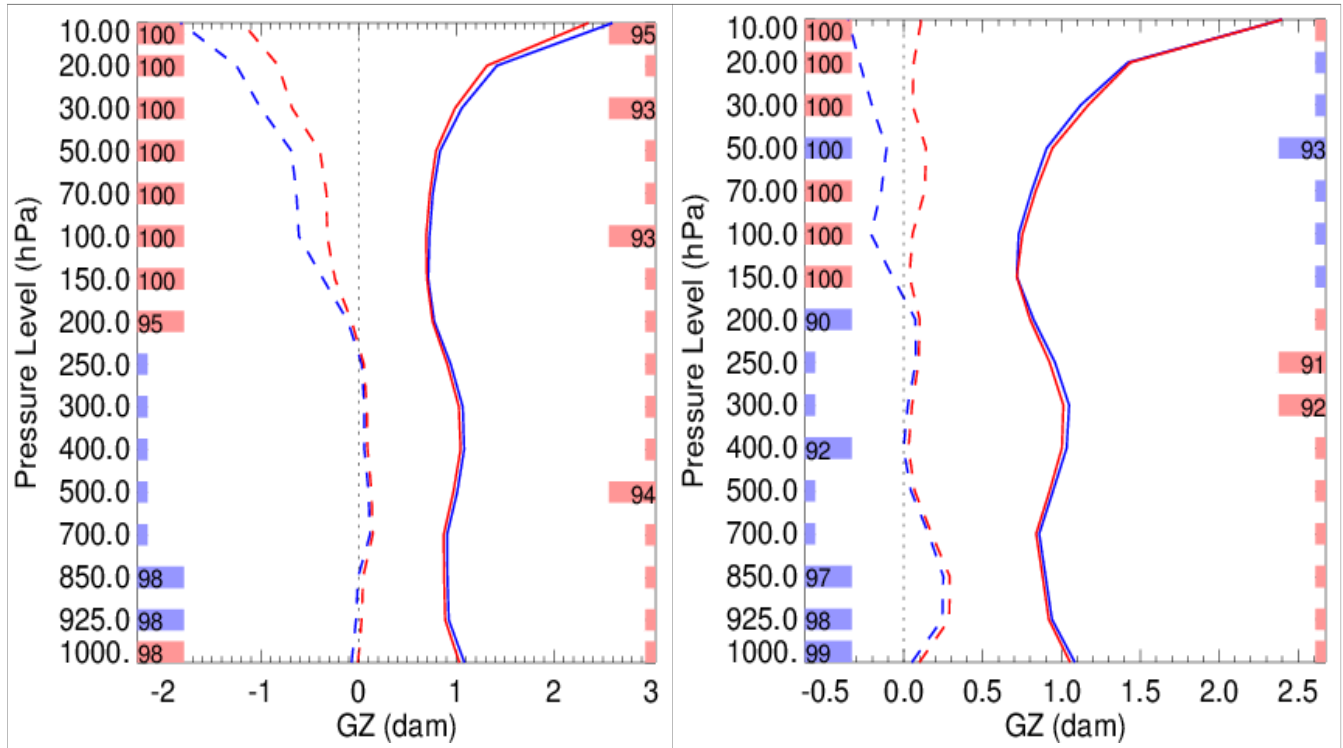


Figure 4: Bias (dashed curves) and error standard deviation (continuous curves) of the final cycle of the CMC Global Deterministic Prediction System with (red) the Yin-Yang grid and (blue) the operational uniform longitude-latitude grid, evaluated from 15th Dec 2014 to 15th Feb 2015 in the North Polar Region. Inference on the differences between the scores of the two systems is performed with a permutation test: significance levels are listed along the vertical axes, on the left for the bias and on the right for the error standard deviation, with red shading indicating an improvement of the Yin-Yang versus lat-lon grid. As expected for these short lead times (24h), verification against own analysis (right panel) shows better statistics than verification against ERA-interim (left panel). Images are courtesy of Stephane Laroche and C. Charette (RPN/MRD/EC).

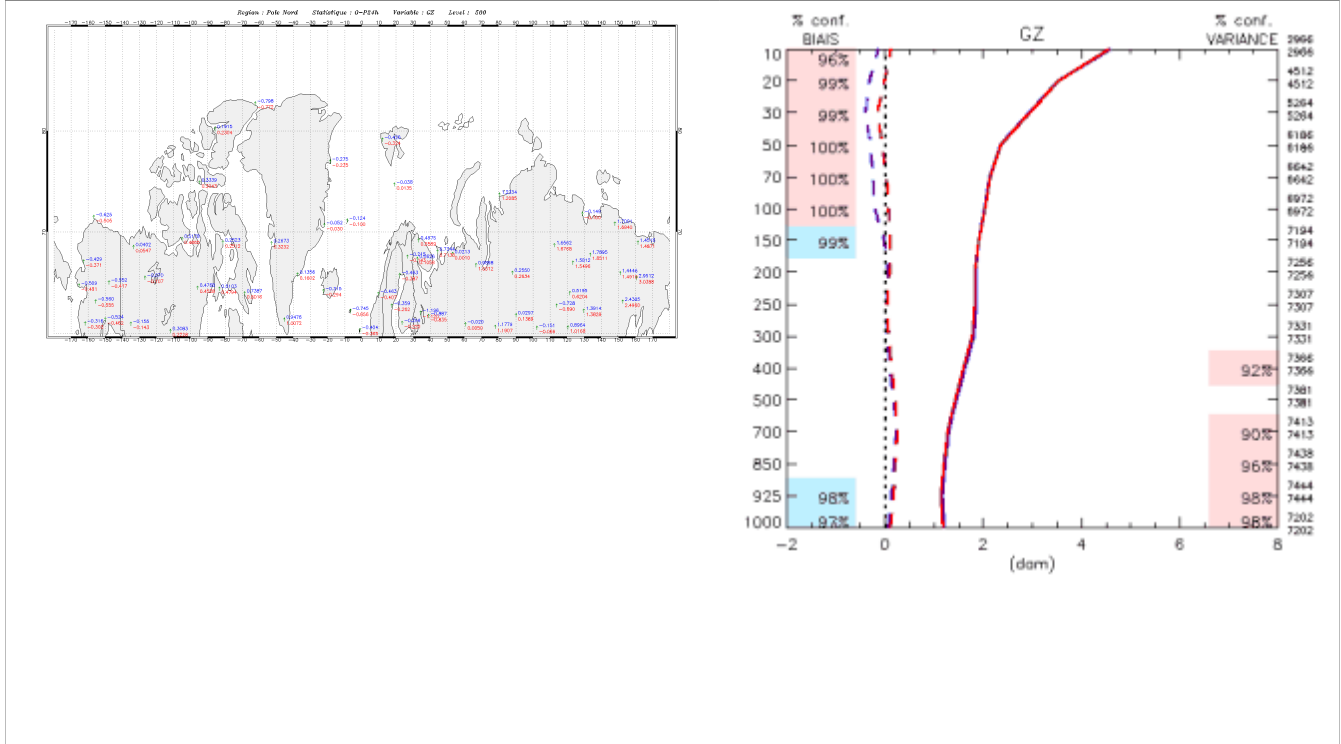


Figure 5: Right panel: bias (dashed curves) and error standard deviation (continuous curves) of the final cycle of the CMC Global Deterministic Prediction System with (red) the Yin-Yang grid and (blue) the operational uniform longitude-latitude grid, against radiosondes, evaluated from 15th Dec 2014 to 1st March 2015 in the North Polar Region (left panel). Inference on the difference between the scores of the two systems is performed with a standard t-test for the bias and F-test for the error standard deviation, for paired samples; significance levels are listed along the vertical axes, on the left for the bias and on the right for the error standard deviation, with red shading indicating an improvement of the Yin-Yang versus lat-lon grid. Images are courtesy of Stephane Laroche and Michel Roch (RPN/MRD/EC).

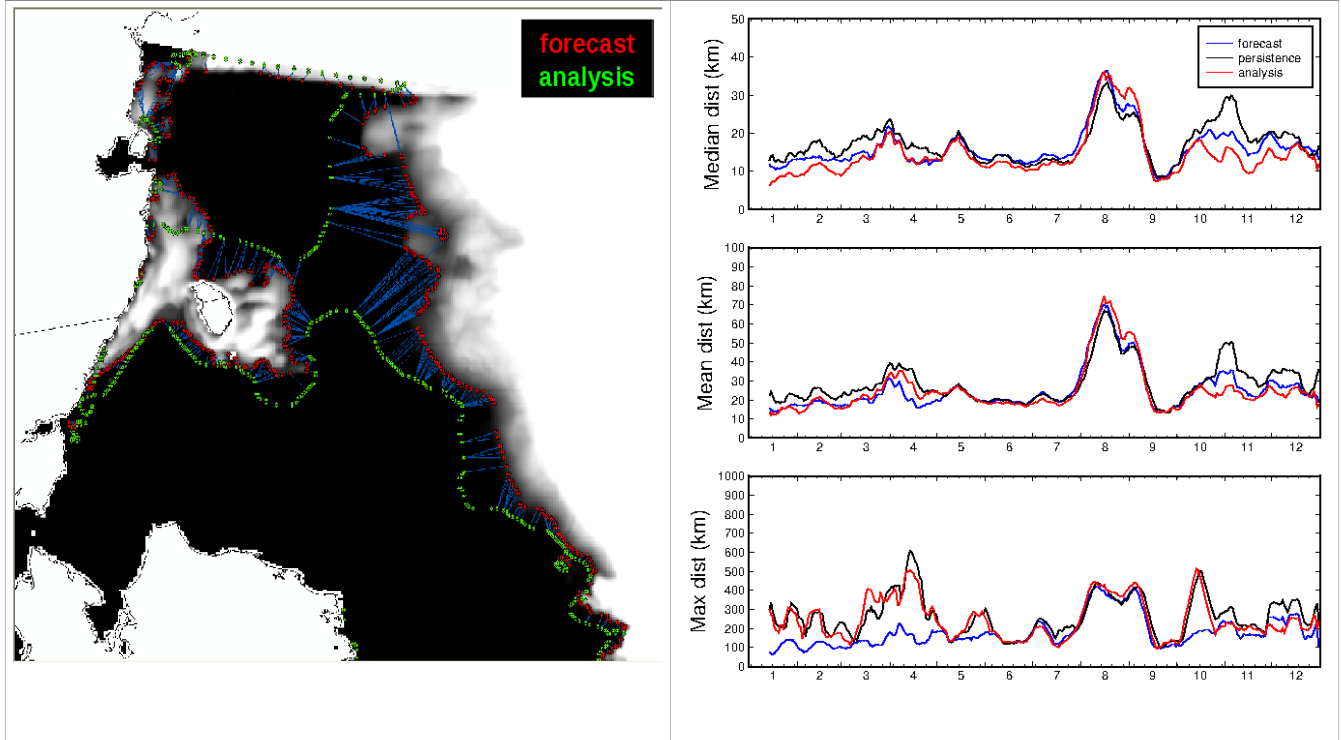


Figure 6: Distance to ice edge. Left: graphical display of distances between forecast and observed ice edge, which are used to evaluate the (partial and modified) Hausdorff distances (analysis and image is courtesy of Angela Cheng, Canadian Ice Service / Environment Canada). Right: verification of the RIPS sea-ice model versus IMS sea-ice observations for the entire 2011 year, by using the median, mean and max distance to ice edge (analysis and image are courtesy of J.-F. Lemieux, RPN/ MRD / Environment Canada). These distances belong to the family of the Hausdorff metrics.

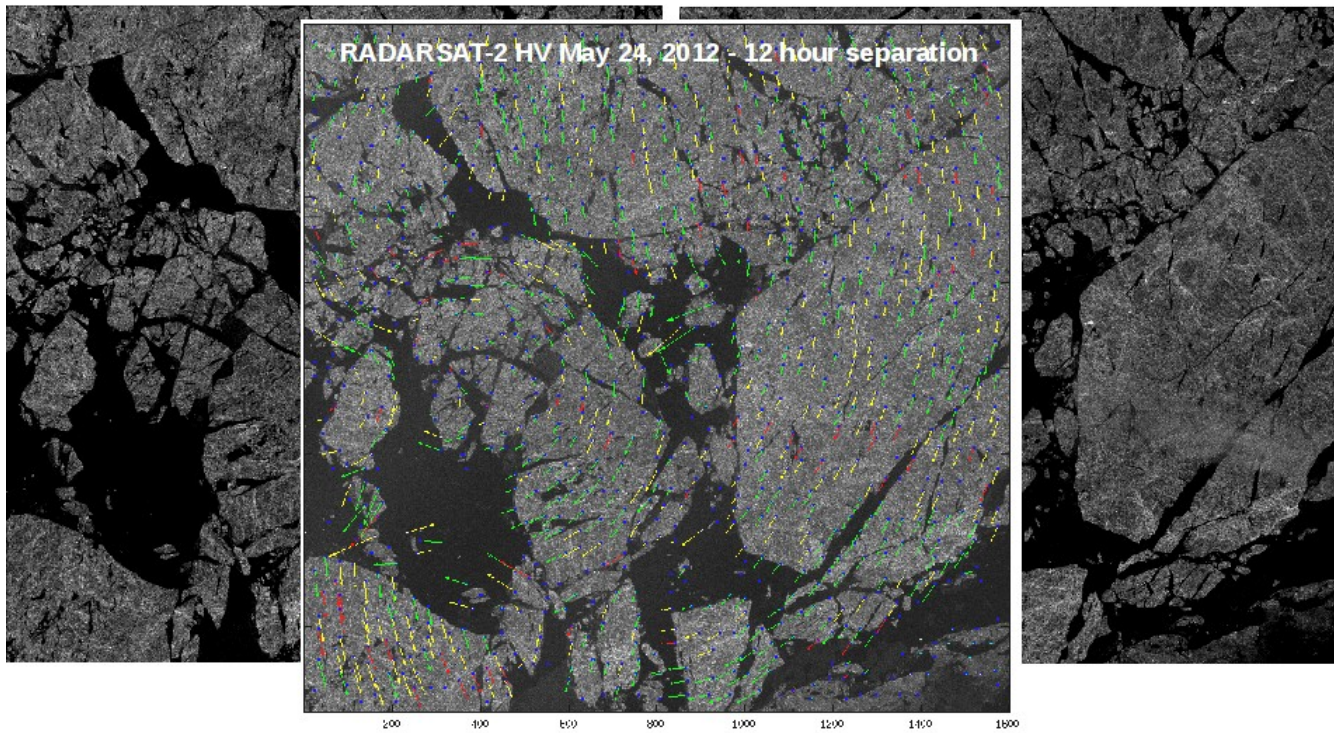


Figure 7: Sea-ice drift field detected by the CIS automated sea-ice tracking system (Komarov and Barber, 2014) from two RADARSAT cross-polarization (HV) Synthetic Aperture Radar images with 12 hours separation, on the 24th of May 2012. Images are courtesy of Angela Cheng, Canadian Ice Service / Environment Canada.