

Chapter 2

Background

Meteorologists create high-quality weather forecasts by synthesizing information from an array of observations and guidance provided by numerical weather prediction (NWP) models. In order for NWP model guidance to be interpretable and useful, it must first be post-processed. Post-processing in general transforms raw model output into a form that is more easily interpreted by the forecaster. Basic post-processing involves interpolating NWP model output from the original grid to other coordinate systems, such as constant pressure levels or height above ground, and calculating derived quantities and diagnostic variables from the fundamental prognostic variables. Feature identification catalogs different areas of interest for particular forecasting tasks. Statistical post-processing combines NWP model output with outside observations and other data sources to produce a calibrated forecast product. This chapter describes these different post-processing methods in more detail and discusses the merits and limitations of different approaches. The ingredients and forecasting approaches for hail and solar irradiance are then examined.

2.1 Convection-Allowing Model Ensembles

As computational speed and storage capacities have increased in the past 20 years, research groups and operational weather forecast centers have run

real-time NWP models capable of explicitly representing deep, moist convection (Kain et al. 2008). NWP models with grid spacing larger than 4 km must parameterize deep convection in order to capture the thermodynamic and precipitation effects from convective overturning properly. At coarser grid spacings, convection occurs on a slower time scale than observed, resulting in convective structures, heat and moisture fluxes, and precipitation amounts being represented incorrectly (Weisman et al. 1997). While not all convective processes are adequately resolved between 1 and 4 km, individual updrafts and their associated heat and moisture fluxes can be represented with a reasonable degree of accuracy. Because some convective processes are not fully resolved, these models are called Convection Allowing Models (CAMs). Even with their imperfect representation of convection, CAMs have shown skill over mesoscale NWP models in precipitation forecasting and in forecasting convective evolution and morphology (Weisman et al. 2008). However, errors in the initial conditions and model physics lead to spatial errors in storm placement and timing errors in convective initiation and storm evolution. In order to account for the uncertainty associated with these errors, modelers have developed CAM ensembles that perturb the initial and boundary conditions, physics parameterizations, and the dynamical cores to create a set of realizations that capture the range of possible convective solutions for a given day. CAM ensembles of many different configurations have run in real time as part of the NOAA Hazardous Weather Testbed Spring Experiment since 2007 (Clark et al. 2012a). The National Weather Service is already running deterministic CAMs operationally and is planning to deploy CAM ensembles in the near future (Benjamin 2014).

While CAMs cannot directly resolve the severe hazards associated with thunderstorms, which include tornadoes, hail, high winds, and flash floods, diagnostic information extracted from individual storms has shown skill in inferring the potential for these hazards. Because storms can greatly evolve in structure and intensity on the order of minutes, hourly instantaneous snapshots of CAM output will likely miss the time when the storm is at its greatest intensity. More frequent model output may capture these extremes, but requires more storage space. A useful compromise is to track the maximum value of a quantity at a grid point over a given time period and output that field every hour. These hourly maximum fields can provide a proxy for both storm track and intensity (Kain et al. 2010). For severe weather forecasting, updraft helicity, the integrated product of vertical velocity and vertical vorticity, between 2 and 5 km is a good proxy for strong, rotating updrafts (Kain et al. 2008) and shows some correlation with tornado path length (Clark et al. 2013). Other hourly maximum fields include updraft and downdraft speeds, radar reflectivity at the -10C level and column-integrated graupel.

Products derived from the hourly maximum fields can help diagnose severe weather likelihood and intensity. The length and direction of the tracks provides an estimate of storm speed and motion. Reliable probabilistic guidance for severe weather can be derived from hourly maximum fields by selecting a threshold, marking grid points where the threshold was exceeded within a given radius, and then applying a Gaussian smoother to the surrogate severe report grid (Sobash et al. 2011). These probabilities can be generated for both deterministic and ensemble configurations (Fig. 2.1) and are very computationally efficient to generate. Varying the standard deviation of the Gaussian smoother

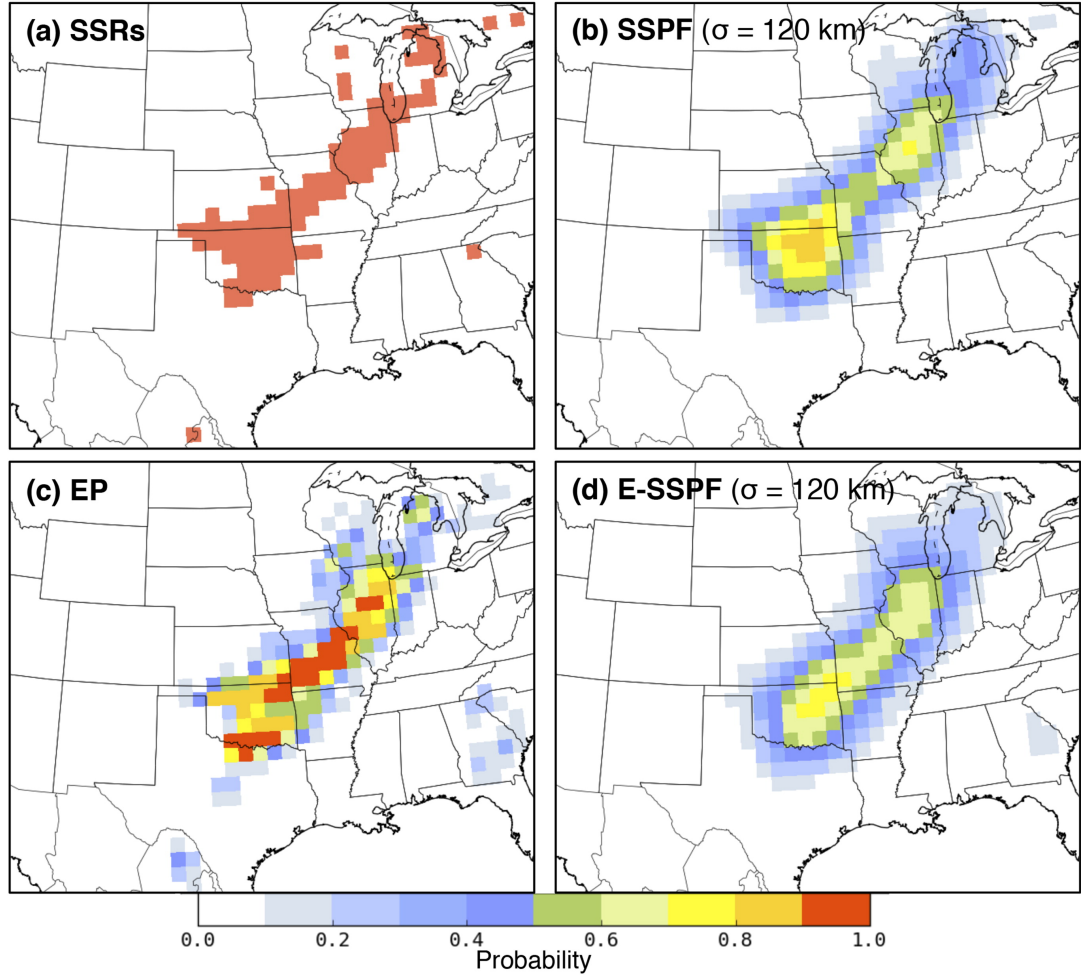


Figure 2.1: (a) Storm Surrogate Reports on an 80 km grid. (b) Storm Surrogate Probability Forecast from a single ensemble member. (c) Ensemble probability (mean of (a)) on coarse grid. (d) Ensemble Storm Surrogate Probability Forecast. Figure from Sobash et al. (2016).

can allow the user to calibrate the storm surrogate probabilities based on the severe weather report of choice (Sobash et al. 2016).

Storm surrogate products have shown usefulness as a forecast guidance product but have some inherent limitations. The current storm surrogate fields do not directly predict the occurrence of severe weather. The values of the storm surrogate products are model core, grid spacing, and parameterization scheme dependent, which can make them more challenging for forecasters to process when examining model output from multiple ensemble prediction systems. Most of the storm surrogate products are only indirectly verifiable in that the surrogate variables are not observed but can be correlated with severe reports. A forecast product that directly forecasts the chance or occurrence of a particular severe threat would be more useful if the predictions were well-calibrated and if they were physically consistent with the model forecast.

2.2 Feature Identification and Tracking Methods in Meteorology

Feature-based post-processing techniques are often used for forecasting atmospheric phenomena that occupy a discrete area and move and evolve with time. Commonly tracked features include cyclones (Blend and Schubert 2000), precipitation areas (Davis et al. 2009), thunderstorms (Dixon and Wiener 1993), fronts (Hewson 1998), and jet streams (Limbach et al. 2012). With feature identification and tracking, scientists can catalog the locations, intensities, and durations of features and compare them across space, time, and different datasets. Feature-based datasets generally require much less data to be stored per event, which allows for larger archives given fixed storage amounts. High-impact

weather events tend to be associated with discrete features, so using feature-based analysis for forecasting can reduce the computational and cognitive load for both forecasters and their guidance models. This dissertation using feature identification and tracking on the hail forecasting problem to identify potential hailstorms and extract information about them that can be fed into machine learning models.

This analysis can be done either subjectively by individuals who hand-label each feature or objectively by automated algorithms that apply the same criteria to every event. Subjective feature identification takes advantage of people’s natural pattern recognition skills and is better for capturing complex features and edge cases. However, the process is very labor intensive and time-consuming (Lakshmanan and Smith 2009), and it often requires someone with training and expertise on the given phenomenon. Subjective identification can also be inconsistent with different experts, or even the same expert at different times, analyzing the same feature or similar features differently depending on their experience or fatigue levels (Hewson 1998). Objective, or automated, feature identification tends to be faster and more consistent than humans and can be run in real-time or archival situations. It can be scaled very cheaply across multiple processors or machines, and it can be performed with different settings on the same dataset to ensure greater robustness. On the other hand, automated approaches generally require a lot of up-front labor to develop and often require data to first be quality-controlled and smoothed in order to produce good results. While many automated techniques are available for feature identification, all of them have to be fine-tuned for the needs and challenges of a particular domain before being used operationally.

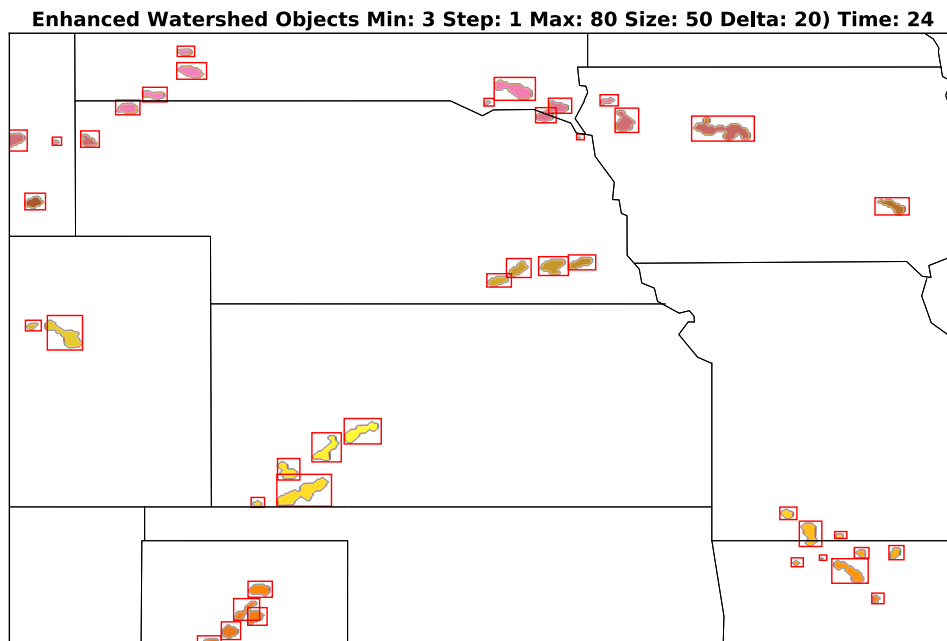


Figure 2.2: An example of features identified in a total column graupel field using the enhanced watershed technique on convection-allowing model output. The features are colored by ID number. The red rectangles show the bounding box around each feature.

Feature identification on a spatial grid typically involves a process of identifying candidate center points, growing regions of influence, merging features, and filtering those that do not meet certain criteria. Simple feature identification methods, such as those used in TITAN (Dixon and Wiener 1993), SCIT (Johnson et al. 1998), and MODE (Davis et al. 2009), look for contiguous areas that exceed a single intensity threshold, but they may capture too many spurious objects if the threshold is set too low or merge objects together that should be considered separate. Objects with maximum values near the threshold may also disappear and reappear due to small fluctuations and would be considered separately by the algorithm. The hysteresis method, which requires each feature to contain at least one point exceeding a higher threshold in addition to all points exceeding a lower threshold, helps filter some spurious objects. The enhanced watershed method (Lakshmanan et al. 2009) grows objects until they reach a specified saliency, or area criterion. This change in criteria makes the method more scale-aware and reduces its sensitivity to the choice of intensity threshold. An example of the features identified by the enhanced watershed method is shown in Fig. 2.2.

Feature tracking methods use some combination of centroid matching and feature cross-correlation. In centroid matching methods, distances are computed from all centroids at one time step to all centroids at another time step. Then features meeting the minimum distance criteria are matched, and those without a matching pair are considered either terminated or new features. The TITAN storm-tracking algorithm (Dixon and Wiener 1993) uses a globally optimal matching algorithm (Munkres 1957) to find the best pairings of storms and to resolve track continuations in the cases of mergers and splits. Han et al. (2009) created an enhanced version of TITAN that first matched objects that

overlapped spatially before matching with a global cost function. Lakshmanan and Smith (2010) evaluated 5 commonly used storm tracking algorithms and devised a new hybrid tracking algorithm that combined the best features from all of the other ones. Limbach et al. (2012) used the overlap of jet stream features in time and space to perform 4-dimensional tracking. The MODE-Time Domain algorithm (Clark et al. 2014) also uses feature overlap to track objects in time.

2.3 Statistical Post-Processing

Statistical models for post-processing NWP model output have evolved within two general frameworks. Perfect-prog, the first framework, fits a statistical model between observed or analyzed variables and observations of a weather feature, such as temperature or precipitation (Klein et al. 1959). The statistical model is then applied to NWP forecasts that assume the model is perfect. Model Output Statistics (MOS), the second framework, fits a statistical model between NWP output at a given time horizon and observations at that time (Glahn and Lowry 1972). Because MOS fits to the NWP output directly, it can correct for systematic biases assuming that the configuration of the NWP model does not change. When NWP model configurations are updated, the MOS equations have to be regenerated after a sufficient number of new model forecasts are made. Perfect-prog models are generally less accurate than a well-tuned MOS model, but they are less sensitive to model configuration changes and tend to be more robust over time.

Both perfect-prog and MOS models have traditionally used multiple linear regression in which the input variables are selected through an iterative screening process. Hundreds of potential input variables are usually available, but the strength of the correlation with the predicted value may be low, or they may be highly correlated with other input variables (Glahn and Lowry 1972). Forward screening selection addresses this problem by fitting linear regression models to all input variables and selecting the model that produces the largest reduction in variance, then finding a second variable that provides the largest reduction in variance when combined with the first, and so on until some stopping criteria is met in terms of number of variables or minimal reduction in variance (Glahn and Lowry 1972). The method produces a set of strong predictors that also minimize cross-correlation, but also requires fitting thousands to millions of linear models before settling on a final one. MOS can produce deterministic, categorical, or probabilistic predictions that are constrained by transforming the input variables into binary values.

The training and forecast data for MOS models need to exhibit as much stationarity as possible in order to maximize predictive skill. MOS equations are generally trained for single sites at each lead time in order to capture local effects closely. For variables that have more skewed distributions, regional aggregation of similar sites helps capture rare events (Lowry and Glahn 1976). While regional models are expected to perform well within their specified region, application of these models on a high-resolution grid can lead to discontinuities at the borders of regions. When developing gridded MOS, Glahn et al. (2009) minimized spatial discontinuities by using a successive correction method (Cressman 1959), spatial smoothing, land-sea masking, and elevation information. The gridded MOS approach does produce effective forecasts, but

it requires a large number of individual regression models for each location and lead time to produce effective forecasts.

2.4 Machine Learning

Modern datasets are constantly increasing in size and complexity, and greater demands are being placed on people to make predictions from these datasets. The assumptions of Gaussian-distributed data and constant variance that underly the simple multiple linear regression used by MOS mean that the approach does not scale well as more examples and predictors are added. Machine learning models, however, are designed to extract information from large, multidimensional, noisy datasets and produce predictions that generalize well. Unlike traditional statistical modeling approaches, which look for empirical distributions that best fit a smaller sample of data as closely as possible, machine learning modeling approaches derive as much information as possible from the data directly and focus on maximizing predictive performance on data independent of what was used to fit the model (Breiman 2001b). Machine learning models make gains in predictive performance by fitting to data in a way that is constrained by model structure and parameters to maximize signal and to reduce noise.

2.4.1 Regularized Linear Models

Regularization techniques allow linear models to fit noisy, high-dimensional data while minimizing the potential for overfitting. Simple linear regression determines its weights by minimizing the mean squared error between the model and observations, which can be done numerically using a gradient descent method. If the input data to the model are noisy such that for a given

set of input values there are multiple output values or contains many input variables with varying degrees of relevance, the least squares fit would either overfit the noise or not converge to an unbiased solution at all. A solution to this problem is to add a bias term to the minimization function that constrains the possible coefficient values.

$$\min_w \frac{1}{2n} \|Xw - y\|_2^2 + \alpha\rho \|w\|_1 + \frac{\alpha(1 - \rho)}{2} \|w\|_2^2 \quad (2.1)$$

Ridge regression (Hoerl and Kennard 1970) addresses this issue by adding a penalty term to the minimization function that is the L2 norm of the coefficient vector (term 3 in Eq. 2.1). The ridge term penalizes large-magnitude coefficients and shifts the optimal solution toward a set of small-magnitude coefficients. Lasso regression (Tibshirani 1996) uses the L1 norm (term 2 in Eq. 2.1) in place of the L2 norm. This minor change results in the potential for a sparse set of coefficients such that some of them are set to 0. This feature of Lasso allows it to do implicit variable selection as part of the optimization process. Unlike ridge, lasso has no analytical solution. The Elastic Net (Zou and Hastie 2005) combines the ridge and lasso terms in the minimization function and balances their effects using a weighting term ρ . The magnitude of their effects is governed by α . By using cross-validation to determine the best ρ and α , a more robust set of coefficients can be found.

2.4.2 Decision Trees

Decision trees are a class of machine learning model that uses a hierarchical set of decision thresholds to recursively partition a complex, multidimensional feature space into many, more uniform subsets (Breiman et al. 1984). A decision

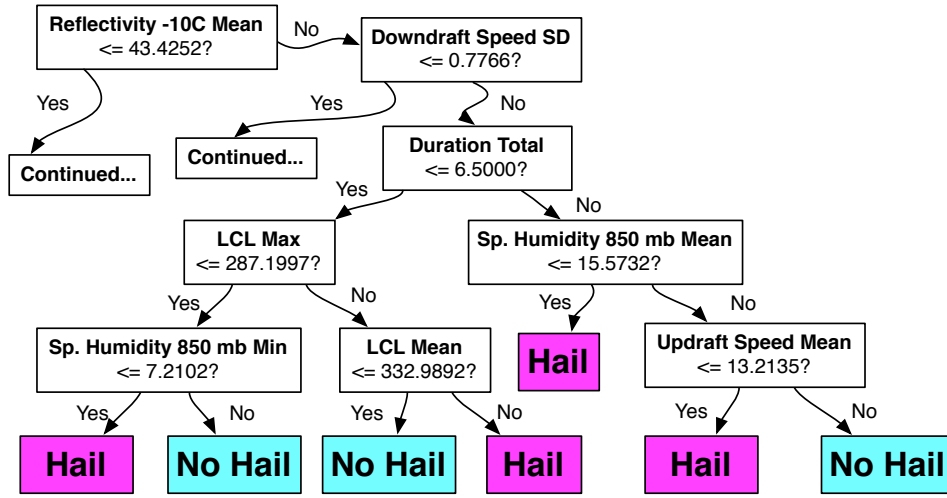


Figure 2.3: An example of a decision tree predicting whether or not hail will occur trained on data from the 2014 Center for Analysis and Prediction of Storms ensemble. See Chapter 4 for more details.

tree consists of binary decision nodes that ask yes-or-no questions about the input features, such as “Does the temperature exceed 20 °C?” The decision nodes eventually branch out to leaf nodes that output a prediction based on the training data that reached them. The prediction can be a class label, probability distribution, or continuous value depending on whether the decision tree is being used for classification or regression. An example of a decision tree can be found in Fig. 2.3.

Decision trees are grown by greedily determining the best input feature and threshold for splitting a given set of data into many more uniform subsets. In the traditional decision tree framework, all input features are evaluated at each node, and candidate splitting thresholds are picked from transitions in labels for the training data. An impurity or error metric, such as entropy or GINI for classification and mean squared error for regression, is calculated for

the current node and for both possible child nodes for every potential feature-threshold combination. The combination with the largest reduction in impurity is selected for inclusion in the tree, and the training data are then split and sent downward where the growth process continues. Decision tree growth stops when a certain level of uniformity, a maximum depth, a minimum number of training cases, or a lack of significant decreases in error is reached. Through the tree-growing process, irrelevant features are not included, and a human-interpretable model is constructed. Because the predictions are only based on the training data labels, decision trees are resistant to noisy and missing input data, but they also do not extrapolate beyond values found in the training set.

Decision trees also have many inherent limitations when used in isolation. When compared with other machine learning models, decision trees often produce less accurate predictions. The models are sensitive to slight changes in the training data, which can have a cascading effect on what input features are included in the tree. Deeper splits in the tree may not be informative if the number of training examples is small, resulting in overfitting (Hastie et al. 2009).

2.4.3 Decision Tree Ensembles

Some of the limitations of decision trees can be ameliorated through different forms of ensembles along with the addition of stochastic processes. Two of the most powerful decision tree ensemble techniques are random forests and gradient boosting trees. Random forests (Breiman 2001a) are unweighted ensembles of randomized decision trees. Each decision tree in the random forest is trained using a bootstrap-resampled dataset, and a random subset of the input variables

are evaluated for inclusion in the tree at each node during the tree growth process. These randomization steps increase the independence of the trees in the ensemble and lead to greater exploration of the feature space (Ho 1998). Evaluating random subsets of the features instead of all possible ones also reduces the computational cost of building a large ensemble. By averaging the predictions from a set of low-bias trees, the variance of the predictions is reduced, so the resulting predictions generally are smooth and have low error. Growing each tree independently also means that the tree growth can be easily parallelized for large datasets. Compared to other complex machine learning methods like neural networks and support vector machines, random forests have few tunable parameters, and a wide range of parameter values will generally produce good performance. Due to their strong performance and relative ease of use, random forests have been increasingly adopted by the members of the weather community for a wide range of nowcasting and forecasting problems, including storm classification (Lakshmanan et al. 2010; Gagne II et al. 2009), convection initiation (Ahijevych et al. 2016), aircraft turbulence (Williams 2014), and hurricane power outages (Nateghi et al. 2014).

Gradient boosting trees (Friedman 2001) are additive, stagewise ensembles of decision trees. In a stagewise ensemble, a series of relatively weak models are fit sequentially to minimize the errors on the training set predictions made by the previous model. In the case of gradient boosting trees, the weak model is a decision tree. The initial tree is fit directly to the training set labels while each additional model is fit to the negative gradient of the loss function. For the mean squared error loss function, the negative gradient is simply the residual of the training label and the prediction from the previous tree. Training examples that have larger residuals will receive higher weights in the fitting process. The

predicted residual from each tree is then added to the sum of the previous tree predictions. Because adding more trees to boosting eventually leads to overfitting, each tree’s residual prediction is multiplied by a learning rate parameter that reduces the magnitude of the residual by a constant value (Hastie et al. 2009). A smaller learning rate parameter requires more trees to achieve the same level of training error but tends to provide lower testing set error. Unlike a random forest, which tends to perform better with large trees, the gradient boosting procedure favors small trees for both faster computation and lower test set errors. The algorithm is not parallelizable but can be computationally efficient if the tree size is limited. Gradient boosting trees were not widely used in the meteorology community until the AMS Solar Energy Prediction Contest, in which the top 4 teams all used some variation of the model for their entry (McGovern et al. 2015).

2.5 Forecast Ingredients

2.5.1 Hail Forecasting

While hail growth and melting is governed by the complex microphysical processes that a hailstone experiences along its trajectory through a storm, hail forecasting methods have relied on a combination of coarse approximations of the expected storm environment and local hail climatology with varying degrees of success (Johns and Doswell III 1992). Most existing hail forecast methods (e.g., Fawbush and Miller 1953; Moore and Pino 1990; Brimelow et al. 2002) are based on information extracted from a sounding representative of the convective environment. All of these models estimate potential updraft strength from the buoyancy between the melting layer and the equilibrium level as the integrated

buoyancy provides an estimate of the maximum potential updraft speed in a given environment. In order to determine the hail size at the ground, the hail size models also include a melting effects component based on the height of the wet-bulb temperature zero level, which is approximately where melting would begin in hail falling through downdraft air. Shallower, cooler, and drier melting layers lead to less melting; and larger hailstones have higher terminal fall speeds, leading to shorter transit times (Johns and Doswell III 1992).

The sounding-based hail forecasting methods have fundamental limitations that limit their skill. The sounding used to assess hail potential should be representative of the storm environment, which can be problematic when observed soundings are generally available only twice a day. This issue can be addressed by either correcting the sounding based on recent surface observations (Moore and Pino 1990) or by utilizing model forecast soundings near the time and location of expected hail. The updraft strength estimated from the sounding may not be representative of the updrafts that actually produce hail. Updraft speed computed from CAPE is estimated at the top of the updraft and not within the hail growth region, and the largest hail may not come from the strongest part of the updraft because the hail embryos may be lofted out of the hail growth region if the speed is too high (Johns and Doswell III 1992). In storms with tilted updrafts, hailstones may have significant horizontal motions in their growth trajectories (Nelson 1983), leading to additional growing time.

Some of the issues with purely sounding-based hail diagnostics were addressed by feeding sounding information into a 1-dimensional hail growth model. This approach, called HAILCAST (Brimelow et al. 2002), creates an ensemble of updrafts based on perturbations of the temperature and moisture profile. A parameter called the Energy Shear Index governs the predicted lifetime of the

updraft and the amount of entrainment that will reduce updraft speed. Hail embryos are then released into the updraft and grow until they can no longer be sustained by the updraft or the updraft collapses. A bulk melting scheme is then applied to approximate the hailstone size at the ground. Jewell and Brimelow (2009) found that HAILCAST generally provides a reliable forecast of hail size, especially in comparison to other sounding based methods. Adams-Selin et al. (2014) has incorporated a variation of HAILCAST directly into the WRF model that uses modeled vertical velocities in place of those estimated from a sounding profile.

2.5.2 Solar Irradiance Forecasting

With the rapid growth of electricity generation from solar energy, there is a greatly increased interest in forecasting solar irradiance at the surface. Solar irradiance, also known as global horizontal irradiance (GHI), is defined as the amount of radiative energy from the sun striking a flat surface of a fixed area over a fixed time period and is generally recorded in units of W m^{-2} . GHI can be decomposed into direct normal irradiance (DNI), the radiation coming directly from the sun, and diffuse horizontal irradiance (DHI), the total radiation reflected by the sky and clouds (Eq. 2.2).

$$\text{GHI} = \text{DHI} + \text{DNI} \cos(\theta_z) \quad (2.2)$$

DNI is multiplied by the solar zenith angle to determine the component striking a horizontal surface. The amount of solar irradiance at a particular location and time is subject to factors with varying degrees of predictability from

directly computable based on simple geometry to highly uncertain. The top-of-atmosphere irradiance depends on the distance between Earth and the sun, which varies cyclically based on Earth’s orbit. Diurnal effects are determined by calculating the solar zenith and azimuth angles for a location and time. At higher zenith angles, solar radiation has to travel a longer distance through the atmosphere, leading to a larger DHI component. Clouds absorb and reflect solar irradiance to varying degrees depending on their thickness, height, composition, and position relative to the sun. Aerosols directly absorb a fraction of solar radiation and can indirectly lead to the formation of clouds (Jimenez et al. 2016).

The best type of forecast guidance for solar irradiance forecasting depends heavily on the lead time (Diagne et al. 2013). On timescales of a few minutes to an hour, persistence and autoregressive models are hard to beat. From 1 to 6 hours, statistical models and advection of clouds from satellite data perform well. From 6 hours to multiple days, NWP models combined with a MOS-type correction will outperform any purely statistical or extrapolation based methods (Diagne et al. 2013; Lorenz et al. 2009). For MOS-like systems, the modelers either spatially average and bias-correct solar irradiance before feeding it into a simulated PV system (Lorenz et al. 2009), or they predict solar power output directly using an archive of PV power output (Zamo et al. 2014). Lorenz et al. (2009) found that spatial averaging of NWP solar irradiance helped improve predictions by better accounting for cloud effects. Zamo et al. (2014) tested a wide range of machine learning models on predicting PV power output from a group of plants in France, and found that random forests produced the lowest errors and outperformed gradient boosting, linear regression, and support vector

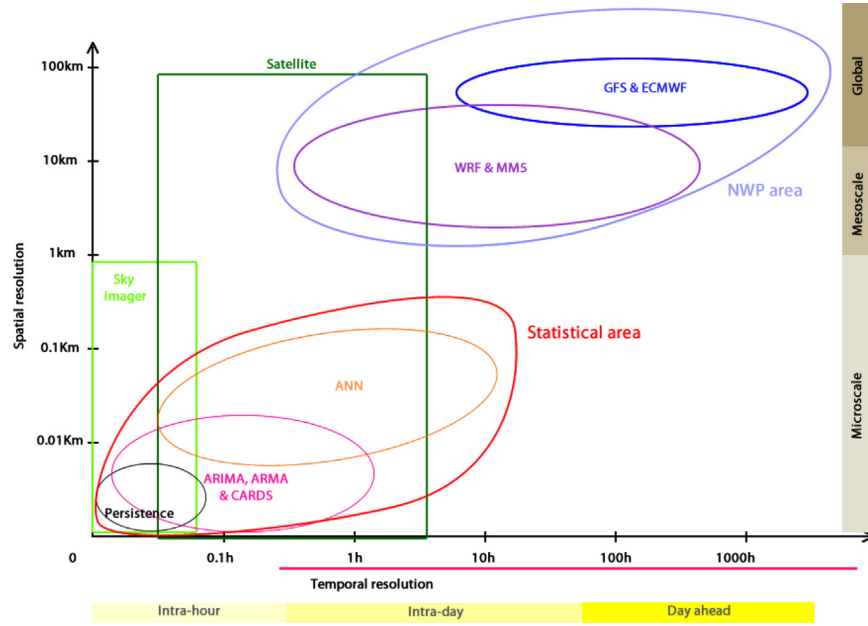


Figure 2.4: Appropriate solar irradiance forecasting models for different time and spatial scales from Diagne et al. (2013).

machines even after having them all optimized through grid search and cross-validation. While a wide range of NWP variables were provided, solar irradiance and sun angle variables dominated the variable importance distribution.

2.6 Forecast Evaluation

The goodness of a forecast is a function of its consistency, value, and quality (Murphy 1993). Consistent forecasts reflect the best judgement of the forecaster and are not skewed to to maximize a verification metric. Forecast value is a function of how much an end user benefits from the information in a forecast. Forecast quality is determined by how well forecasts correspond with observations. Consistency is determined by the forecaster or forecast system and is difficult to assess directly. Value is dependent on the end user and their circumstances and can vary greatly such that a lower quality forecast could potentially have more value. This dissertation focuses on evaluating different aspects of forecast quality. Forecast quality is described by several related attributes that can be expressed as scalar scores (Wilks 2011). Wilks (2011) describes six of the primary aspects of forecast quality that manifest themselves through different evaluation techniques. Accuracy is to the correspondence between individual forecasts and observations. Systematic bias is the difference between the mean forecast and mean observation. Reliability, or conditional bias, measures the distribution of observations conditioned on a certain forecast value. Resolution refers to the differences in observational distributions for different forecast values while discrimination refers to differences in forecast distributions for different

Table 2.1: Example of a binary contingency table.

		Observed	
		Yes	No
Forecast	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

observed values. Finally, sharpness describes the width of the distribution of forecasts, in particular the amount of deviation from climatology.

Most types of forecasts can be reduced to a deterministic prediction of whether or not a certain event will occur. Verification of binary discrete forecasts is performed with a binary contingency table (Table 2.1). The table represents the joint distribution of all possible forecast and observation pairs. Scores describing many properties of the forecasts can be calculated from the contingency table (Table 2.2). Since high impact weather events tend to be rare, there is more interest in predicting the positive events correctly. However, the negative events usually far outweigh positive events in frequency, so statistics that give equal weight to positive and negative events or only focus on negative events, such as Percent Correct and Probability of False Detection, will still return really high scores even if the positive events are poorly forecasted. Critical Success Index (CSI; Gilbert 1884) serves as a better measure of accuracy for rare events by ignoring true negatives entirely. Frequency bias determines if an event is being overforecasted ($\text{bias} > 1$) or underforecasted ($\text{bias} < 1$). False Alarm Ratio (FAR) is a measure of reliability for a positive event while Probability of Detection (POD) measures the ability of the forecast to discriminate between positive and negative events (Wilks 2011).

Probabilistic forecasts of binary events are evaluated using diagrams and scores that examine the different properties of forecast quality at each probability threshold. The accuracy of probability forecasts is evaluated with the Brier

Table 2.2: Verification scores derived from a binary contingency table.

Score Name	Equation
Percent Correct (PC)	$\frac{TP+TN}{TP+TN+FP+FN}$
Critical Success Index (CSI)	$\frac{TP}{TP+FP+FN}$
Probability of Detection (POD)	$\frac{TP}{TP+FN}$
Probability of False Detection (POFD)	$\frac{FP}{TN+FP}$
False Alarm Ratio (FAR)	$\frac{FP}{TP+FP}$
Success Ratio (SR)	$\frac{TP}{TP+FP}$
Frequency Bias (Bias)	$\frac{TP+FP}{TP+FN}$

Score (Brier 1950), which shows the squared difference between probability forecasts and the occurrence of the event. The Brier Score can be decomposed into three terms (Eq. 2.3)

$$BS = \frac{1}{N} \sum_{k=1}^K n_k (p_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^K n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1 - \bar{o}) \quad (2.3)$$

where N is the number of forecasts, K is the number of probability bins, n_k is the number of forecasts in each bin, p_k is the forecast probability, $\bar{o}_k$ is the observed relative frequency for a given probability, and $\bar{o}$ is the climatological relative frequency (Murphy 1973). The first term indicates the reliability of the probabilities as the distance between the forecast probability and observed relative frequency, the second term determines the resolution as the difference between the observed and climatological relative frequencies, and the uncertainty reflects the underlying predictability of the problem. Rare events tend to have low uncertainty. Good probability forecasts minimize the reliability term while maximizing the resolution term. The attributes diagram (Hsu and Murphy 1986) is a graphical representation of the Brier Score decomposition.

The diagram expands on the reliability diagram, which plots the forecast probability versus the observed relative frequency, to include lines demarcating “No Skill” and “No Resolution”. The “No Skill” line marks the points where the reliability term equals the resolution term, which results in a Brier Skill Score of 0 (Hsu and Murphy 1986). The “No Resolution” line indicates probability forecasts that have the same observed relative frequency as the climatological probability. The attributes diagram can be used to assess whether the forecasts for a particular probability threshold are contributing resolution and positive skill. Forecasts with negative Brier Skill Scores can still be useful if the slope of their reliability curve is positive, which indicates the potential for additional calibration at the expense of sharpness (Wilks 2011).

The ability of a probabilistic forecast to discriminate between two outcomes can be assessed using a Relative Operating Characteristic (ROC) diagram (Mason 1982). The diagram is plot of Probability of False Detection on the x-axis versus Probability of Detection on the y-axis. A set of thresholds are chosen to make binary deterministic forecasts from probabilistic on continuous-valued forecasts, and binary contingency tables are constructed for each threshold. The POD and POFD are calculated at each threshold to form a curve. At the lowest threshold, all forecasts are yes forecasts, which results in a POD of 1 and a POFD of 0. At the highest threshold, the opposite is true, resulting in a POD of 0 and POFD of 1. If the POD equals the POFD, then the forecast is no better than a random or uniform forecast. Positive skill over climatology occurs when the POD exceeds the POFD. The area under the ROC Curve (AUC) is a summary metric for the skill across all thresholds with 1 indicating a perfect forecast and 0.5 equal to a random forecast. The ROC curve is insensitive to calibration and the underlying distribution of the event in question, which

makes it a good estimator of potential skill but could lead to poor conclusions if calibration is important (Wilks 2011). Because ROC curves weigh positive and negative events equally, ROC curves may not be the best tool for evaluating rare event forecasts. The performance diagram (Roebber 2009), a variation on the ROC curve that replaces the probability of false detection with the false alarm ratio along the x-axis, is able to display measures of accuracy, bias, reliability, and discrimination all in one diagram and is more sensitive to the ability to predict positive events.

Evaluation of continuous-valued deterministic forecasts can also be framed in a distributions-oriented way. Accuracy is assessed through the mean squared error (MSE) and mean absolute error (MAE). The MSE is the mean of the squared difference between each forecast and observation pair while the MAE is the mean of the absolute value of the difference between each forecast and observation pair. MSE is differentiable for all possible error values, which makes it a popular choice for loss functions (Hastie et al. 2009), but the squared error penalizes large deviations heavily and makes the score more sensitive to outliers. MAE linearly weights errors, so it is viewed as a more robust score. Mean error (ME), or the mean of the differences between forecasts and observations, is a measure of systemic bias. Reliability and discrimination can be assessed by binning the continuous forecasts and examining the marginal distributions of the observations conditioned on the forecasts and forecasts conditioned on the observations, respectively. Sharpness is proportional to the variance of the forecasts.

One major goal of forecast evaluation is determining which forecast system produces the highest quality forecasts. In order to do so, multiple forecasting

systems are evaluated over a large sample of cases using a common set of measures and are ranked. If two forecast systems have similar scores, then their differences and relative rankings may be due to random chance within the sample rather than forecast system design. Statistical hypothesis testing can help determine if two sets of forecasts originate from the same distribution or not. Traditional hypothesis testing requires the assumption of certain properties of the forecast distribution, such as normality. Resampling tests, however, are non-parametric and only require that each item in a sample be independent of the others (Efron and Tibshirani 1994).

The bootstrap is a method for calculating confidence intervals for any arbitrary statistic by repeatedly resampling a dataset with replacement and calculating the statistic on each of the replicate samples. The percentiles of the bootstrap statistic distribution then are used as confidence intervals for that statistic. The width of the bootstrap confidence interval indicates the uncertainty of the statistic calculated on the dataset, which is a function of the variance and sample size of the dataset. The difference between two samples can be inferred as statistically significant if the bootstrap distribution of the difference in sample statistics does not overlap 0. If two bootstrap distributions do not overlap each other, statistical significance can be inferred; but if there is overlap, statistical significance may still be possible depending on the amount of overlap (Cumming and Finch 2005).

If bootstrap confidence intervals computed independently for each forecasting system overlap, then statistically significant differences in performance can still be inferred if the different models forecast the same cases. If the case indices are resampled the same way for each model, then the performance statistics calculated for each bootstrap replicate can be ranked. If one model is ranked

higher than another model for most bootstrap samples, then there is more confidence that the difference in scores is statistically significant even if the actual difference is small. Small but consistent differences in performance are important in applications where many forecasts are used, and forecast errors lead to decisions that incur a certain cost. The cumulative cost reduction of a small but consistent forecast improvement could be very significant, particularly in a domain like electric utility forecasting where power purchasing decisions are made hourly.

Permutation tests can estimate the p-value of the difference in a statistic between two paired sets of data (Efron and Tibshirani 1994). Paired data means that each item in one sample is paired with another item from the other sample, such as forecasts from two different models for the same event. In a permutation test, the difference in test statistics is calculated for the original samples. Then a null distribution of that statistic is computed by randomly shuffling paired items from one sample to the other and recalculating the statistic a large number of times under the null hypothesis that both samples come from the same population. The percentile of the original statistic in the null distribution is its p-value. In cases where the bootstrap confidence intervals of paired data overlap with each other, a permutation test can be used to determine if the difference is statistically significant.