WESTERN REGION TECHNICAL ATTACHMENT
NO. 99-15
August 10, 1999

Model Skill Assessment Resources

Kirby Cook - WRH-SSD

Introduction

Model skill assessment is a very important tool, particularly in the current environment of constantly evolving operational models. Although there is a large pool of information available on the subject, the practical use of this resource is problematic in an operational environment. This technical attachment is intended as a brief review of the basic issues behind model skill assessment as well as a brief summary of some of the resources available on the World-Wide-Web.

Type and Quality of Observations

Because measures of model skill (with respect to traditional validation techniques) are a function of observations as well as model forecasts, care must be taken with the type and quality of observed data used. For the most part, two kinds of observational sets are employed in comparison-driven model validation. Gridded-analysis data sets consisting of grids (usually the same resolution as the forecast data) generated from actual observations, and observations from point sources. Both types have inherent strengths and weaknesses that ultimately affect the quality of skill measures produced from forecast comparisons. Typically, gridded analyses are easy to compare against forecast data (they are usually the model's own initial analysis) and have excellent spatial coverage. This allows for ease of use as well as the ability to draw conclusions regarding the spatial distribution of forecast skill. However, because these data forms are "created" from the actual observations, they may not be truly representative of the nature of the atmosphere. In essence, they may contain errors which ultimately contribute to the end measure of forecast skill within the validation. Point observation data such as surface and upper-air observations are much more representative of the true nature of the atmosphere than gridded analyses. However, they are limited spatially and temporally (such as radiosonde data) and may not resolve certain phenomenon such as mesoscale features. In addition, model data must be converted (interpolated) spatially to such data sets for adequate comparison. Although these data sets may be more difficult to apply to forecast comparisons, if given a large enough sampling they are the most reliable, producing the more realistic measures of model

skill. Finally, regardless of data type, it is important to minimize errors within both the forecast and observational data sets. Quality control must be employed in order to ensure that flawed data is not allowed to corrupt calculated skill measures.

Method of Forecast Comparison

Forecast comparison may be viewed subjectively or quantitatively. In either case the goal is to compare numerical model output to acceptable observations. In the subjective sense, simple, on-the-fly comparisons of model output to satellite imagery, surface and upper-air observations or model analyses (Fig. 1) can provide basic insight into model performance. However, such methods, more often than not, involve human interpretation as the primary mode of comparison. While this can be a reliable tactic in generating an overall measure of model performance, it does not provide a distinct, repeatable calculation.

The objective, or quantitative evaluation of model skill involves the direct comparison of model data and observations. This comparison requires that both data types exist in a common format. Ultimately model forecasts and observations must be linked spatially and temporally (e.g. they must be valid at the right place at the right time). This is simple if the observed data set consists of the model's own gridded analysis. Because both the analysis and prognostic fields share the same grid, forecasted values can be compared directly to the model initial fields at each grid-point and time. However, this comparison is problematic when the observed data consists of point values or an analysis on a different grid. In such situations, the model forecasts must be converted spatially to the observation locations (or vice versa). There several ways of accomplishing this. One such scheme is to simply select the nearest forecast grid-point for comparison to the observed value (Fig. 2). A second option is to estimate forecasted values at observation locations from data at surrounding grid-points (Fig. 3). This estimation can be accomplished using methods such as bilinear interpolation or a higher order scheme such as bicubic spline (Press et. al., 1992). Once the forecasted data has been converted to the observation location, it may be compared to observed data for the generation of skill measures.

Statistical Measures of Model Skill (Objective/Quantitative Comparisons)

The generation of statistical measures of model skill is driven by the need for an objective and succinct description of model performance. There are many different measures, both traditional and non-traditional, each with its own strengths and weaknesses. For the most part, these measures are products of direct comparison of forecast and observation, and can represent values summed spatially or temporally. Below is a brief summary of the most common traditional measures used in numerical forecast evaluation. A more detailed description of statistical measures and their use in atmospheric science is offered by Murphy and Katz (1985).

Standard deviation (SD) is the measure of dispersion from the mean of a particular parameter as illustrated by the equation:

$$SD = \left[\frac{1}{N-1} \sum_{n=1}^{N} \left(x_n - \bar{x} \right)^2 \right]^{1/2}$$

where N is the size of the sample and x_n and $\bar{x}$ are the sample value and mean value of the parameter being measured. A large standard deviation indicates large dispersion from the mean. In terms of error measure, standard deviation is typically used to measure the extent that forecast error differs from location to location from the mean.

Bias error (BE) is a measure of the tendency of a model to under forecast or over forecast a parameter and is defined by the equation:

$$BE(x) = \frac{1}{N} \sum_{n=1}^{N} \left(x^f - x^o \right)$$

where N is the total number of forecast comparisons and f and o denote forecast and observed values respectively. A positive bias error indicates a tendency to over predict a variable while a negative bias error implies a tendency to under predict a variable.

Root-mean-squared error (RMSE) is the square-root of the average of the individual squared differences between forecast and observation. RMSE is defined by the equation:

$$RMSE(x) = \left[\frac{1}{N} \sum_{n-1}^{N} \left(x^f - x^o \right)^2 \right]^{\frac{1}{2}}$$

where, again N is the total number of forecast comparisons and f and o denote forecasted and observed values respectively. RMSE represents the typical size of forecast error, with values equaling or near zero indicating perfect or near perfect forecasts. The squared difference term places more weight on large discrepancies between forecast and observation.

Mean-absolute error (MAE) is the average of the absolute value of the difference between forecast and observation as shown by the equation:

$$MAE(x) = \frac{1}{N} \sum_{n=1}^{N} \left| x^f - x^o \right|$$

MAE values near or equal to zero indicate perfect or near perfect forecasts. This measure is not as heavily weighted towards large differences in forecast comparisons as with root-mean-squared error.

Equitable-threat score (ETS) is a measure of categorical forecast skill as defined by the equation:

$$ETS = \frac{H - C}{F + O - H - C}$$

where H is the number of forecast hits for a particular event, C is the chance of a

random correct forecast, F is the number of forecasts for the event, and O is the number of observed occurrences. ETS values may range from 1.0 to negative values, where a score of one represents a perfect forecast and a score of 0.0 indicates that the skill of the forecast is equal to that of chance.

Online Resources

There are several groups devoted to assessing model performance on a real-time or daily basis. This section will focus on five sites produced at national, regional and university centers throughout the United States. The National Centers for Environmental Predition (NCEP) Evironmental Modeling Center (EMC) Mesoscale Modeling Branch (MMB), Global Modeling Branch (GMB), and Hydrological Prediction Center (HPC) produce measures of model performance over domains covering the Northern Hemisphere, the Continental United States (CONUS) as well as regional subsets. On a regional scale, the Western Region Headquarters, Scientific Services Division of the National Weather Service (WRH) and the University of Washington, Department of Atmospheric Sciences (UW) are producing measures of model skill over the Western United States and Pacific Northwest domains respectively. Each program uses the World Wide Web as the primary mode of distribution of model performance data/information. Table 1 is a listing of each group's URL, while Table 2 displays each location's general characteristics. While nearly all groups produce measures of skill for the Eta, NGM (except WRH), MRF, and AVN forecast systems, the HPC and UW sites provide additional evaluations of other operationally available models (NOGAPS, ECMWF, UKMET and the Canadian Meteorological Center GEM model). All of the above evaluation efforts update daily or several times weekly, depending upon the availability of observations. The WRH and UW efforts make subjective model forecast comparisons against point observations (raob and sfc) using bilinear and bicubic spline interpolation schemes (Press et al. 1992) respectively. The University of Washington site utilizes additional data from satellite (cloud track/water vapor winds, scatterometer winds, and precipitable water), ACARS, ship and buoy observations. These comparisons are made on a real-time basis, updating with the availability of model and observational data. The MMB and GMB comparisons are made using both surface and upper-air observations (bilinear interpolation and nearest gridpoint), gridded data (model analyses and Stage IV gridded precipitation analyses) and satellite derived moisture products (GMB). HPC evaluations are made using gridded model analyses only. Comparisons produced by the EMC modeling groups, WRH and the UW are generated from forecasts stemming from both 0000 and 1200 UTC initialized model runs, while the HPC assessments of model skill are created from 0000 UTC initialized model forecasts only. Forecasted parameters evaluated at each site include temperature, dewpoint, relative humidity, component wind, geopotential height, precipitation (MMB,GMB, and HPC), precipitable water (GMB, UW), and sea-level pressure (see Table 2 for a more detailed distribution). The prevalent statistical measure used by all sites is bias error, however most groups utilize the additional measures of root-mean-squared error, threat-score, mean error, mean-absolute error, and standard deviation. These measures are computed for single forecasts (MMB, GMB, WRH, and UW), 5 day (HPC), 7day (MMB and WRH), 10 day (HPC) and monthly

(MMB, GMB, HPC, and WRH) periods. It is important to note that none of these on-line resources is produced operationally, and that products may not be available in a consistent or timely manner. Their use is intended solely as a supplemental source of information on model performance.

Acknowledgments

The author would like to thank Brett Newkirk (Dept. of Atmospheric Sciences, University of Washington), Dr. Suranjana Saha (Global Modeling Branch, NCEP/EMC), Hua-lu Pan (Global Modeling Branch, NCEP/EMC), Peter Caplan (Global Modeling Branch, NCEP/EMC) and Geoff Dimego (Mesoscale Modeling Branch, NCEP/EMC) for their correspondence and permission to document their on-line resources.

References

Murphy, A. H., and H. Dane, 1992: Forecast Evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Westview Press, 379-438.

Press, W. H., S. A. Teukolski, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran (The Art of Scientific Computing)*. Cambridge University Press, 963 pp.

Table 1. Groups producing daily or realtime measures of model performance on the world-wide-web.

NCEP/ EMC Mesoscale Modeling Branch (**MMB**)	http://ftp.ncep.noaa.gov:8000/research/meso.verf.html
NCEP/EMC Global Modeling Branch (**GMB**)	http://sgi62.wwb.noaa.gov:8080/RTPUB/
NCEP, Hydrological Prediction Center (**HPC**)	http://www.hpc.ncep.noaa.goc/html/hpcframes.html
Western Region Headquarters, Scientific Services Division (**WRH**)	http://www.wrh.noaa.gov/wrhq/DIAGNOSTICS/diag.html
University of Washington, Department of Atmospheric Sciences (**UW**)	http://www.atmos.washington.edu/~bnewkirk/

Table 2. Description of five sites producing model skill assessment on the world-wide-web.

Location	Models	Domain	Observations Used	Comparison Type	Comparison Times (UTC)	Parameters	Statistical Measures	Period of Measures	Updated	Comparison Method
NWS/WR	Eta, MRF/AVN	Western U.S.	uair, sfc	objective	0000 & 1200	hght, tmpc, dwpc, RH, component-wind	BE	single forecast; 7-day and monthly averages	real-time/daily	bilinear interpolation
NCEP/EMC Mesoscale Modeling Group	Eta, MRF/AVN, NGM	CONUS, Regional	grid, raob, sfc	subjective, objective	0000 & 1200	hght, tmpc, RH, component-wind, ppt	BE, RMSE, TS	single forecast; 7-day and monthly averages	ppt daily	bilinear interpolation
NCEP/EMC Global Modeling Group	MRF, ECMWF, UKMET, NOGAPS	Global, Regional	grid, raob, sfc, satellite	subjective, objective	0000 & 1200	hght, tmpc, component-wind, wind spd, sfc pressure, moisture,	BE, RMSE	single forecast, monthly average	daily	nearest gridpoint, bilinear interpolation
NCEP/HPC	Eta, MRF/AVN, NGM, NOGAPS, ECMWF, UKMET,	Northern Hemisphere, Conus	model analyses	subjective, objective	0000	hght, tmpc, RH, ppt	BE, RMSE, TS	single forecast; 5-day, 10-day and monthly averages	real-time/daily	none
U. of Wash.	Eta, MRF/AVN, NGM, NOGAPS, CMC	Pacific N.W., N. Pacific	uair, sfc, sat, ACARS, ship, buoy	objective, subjective	0000 & 1200	tmpc, component wind, slp, PW	RMSE, ME, AME, SDE, ASDE	single	real-time/daily	bicubic interpolation

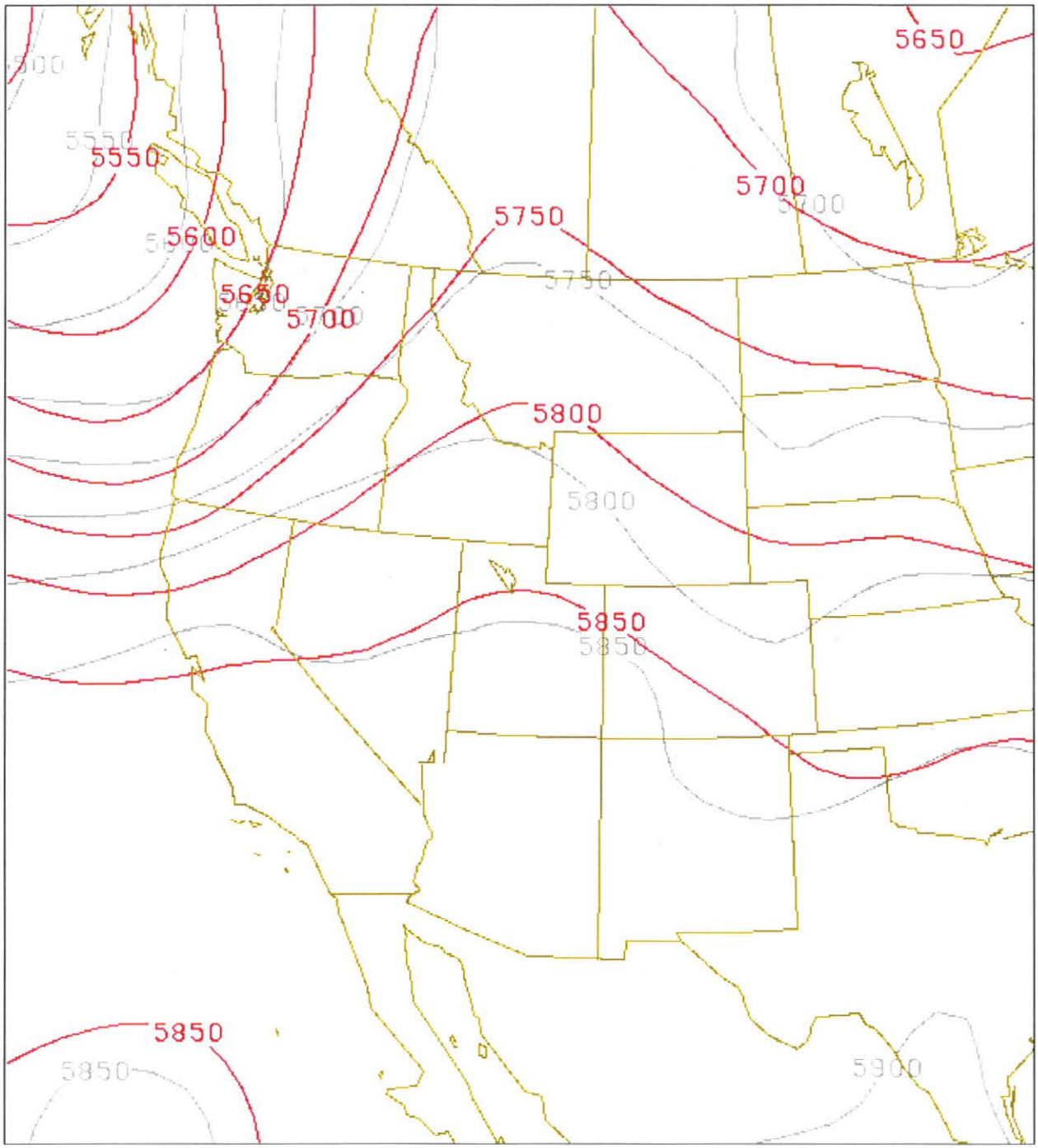

Figure 1. Subjective comparison of Eta (grid 211) 24–h geopotential height forecast against its own analysis valid on 990618, 0000 UTC. The forecast is shown in red while the analysis is shown in grey.

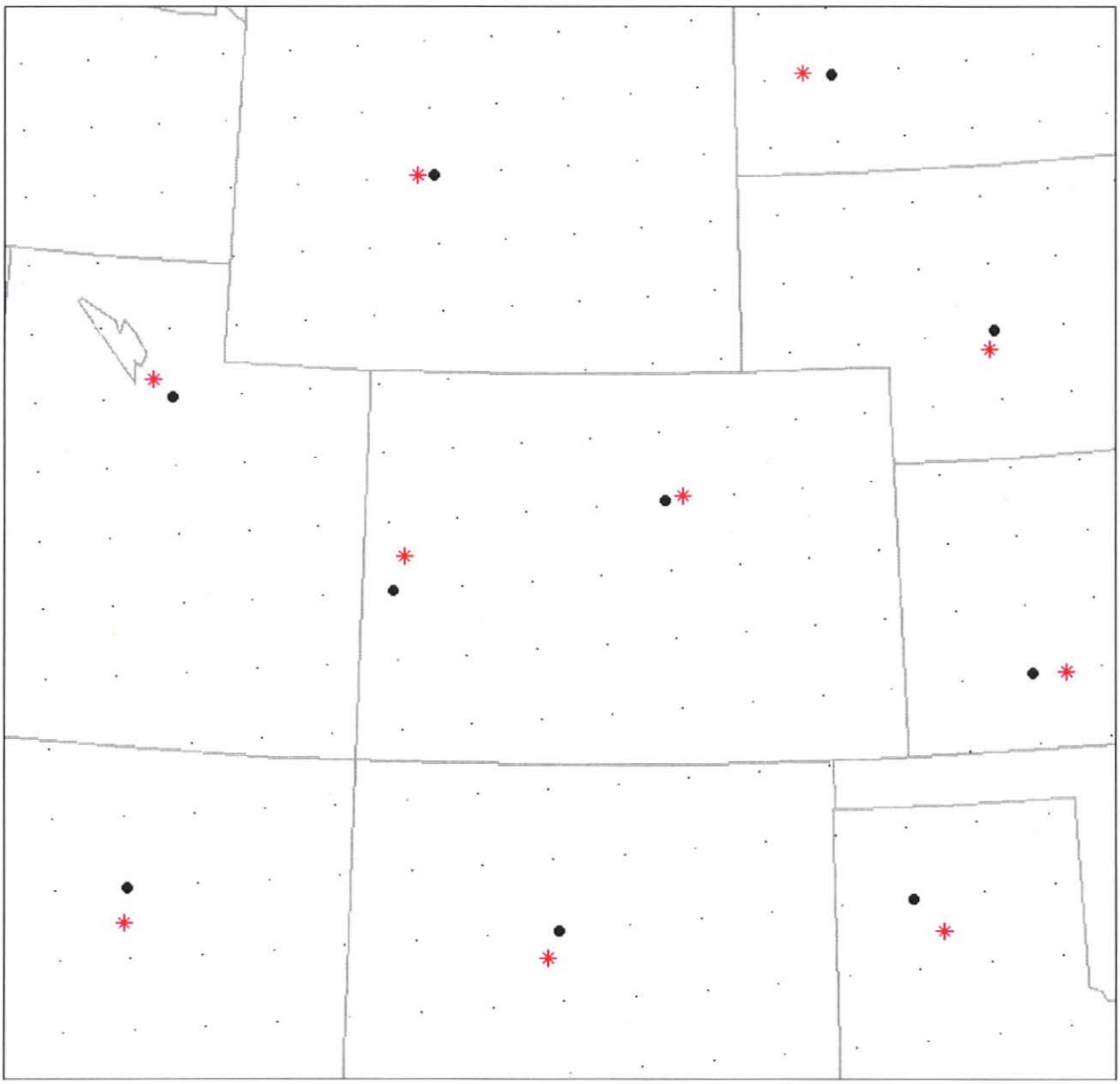

Figure 2 Eta gridpoints (grid 211) with respect to sounding locations (red asterisks) in the Intermountain Region of the United States. Dark black dots denote the nearest gridpoint to each upper-air site.

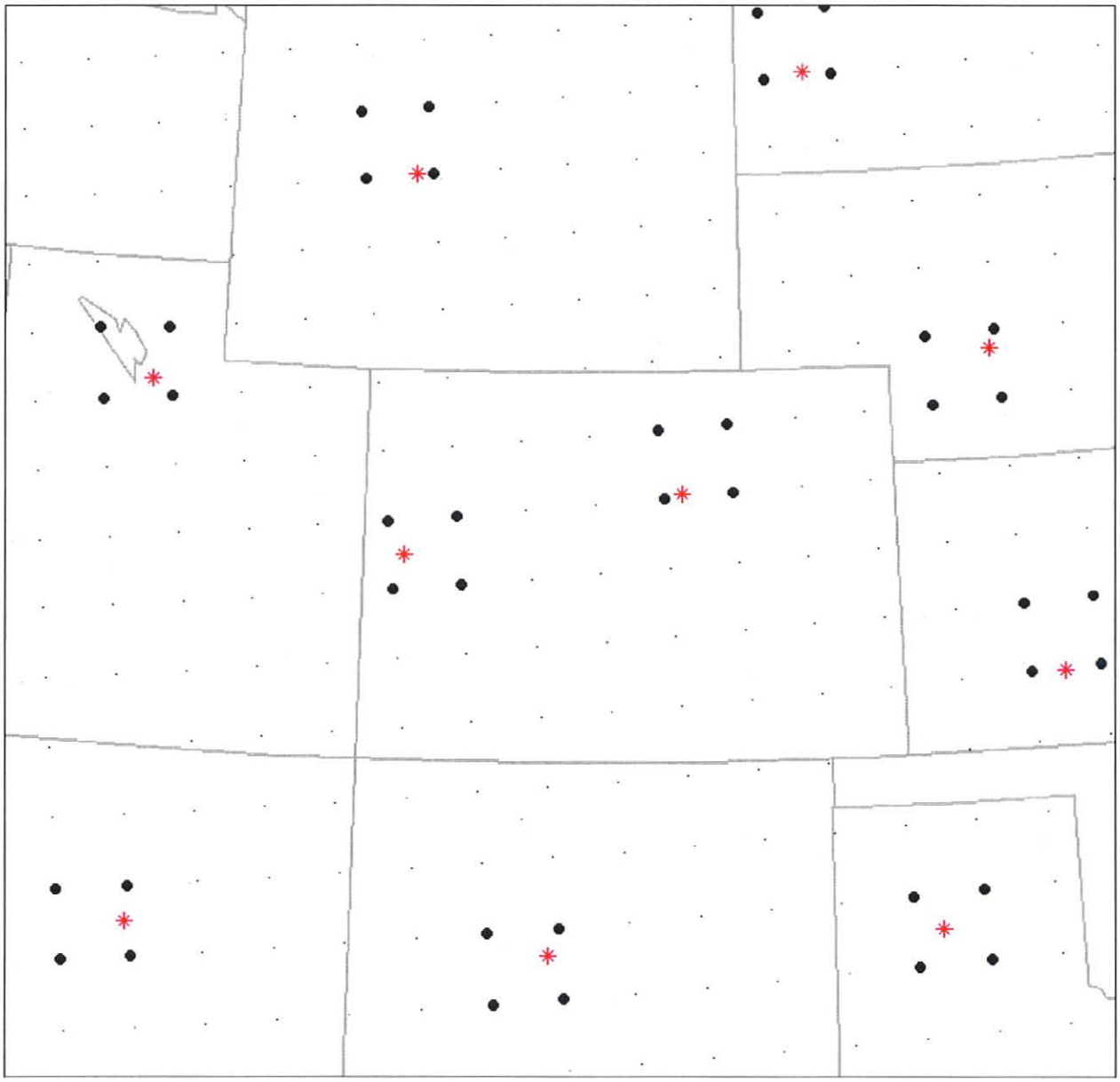

Figure 3. Eta model gridpoints (grid 211) with respect to sounding locations (red asterisks) in the Intermountain region of the United States. Dark black dots denote the 4 nearest grid points to upper-air sites.