2.1 TYPICAL DAY METEOROLOGICAL DATA IN SUPPORT OF ATD MODELING

George D. Modica[1], Steve Lowe, Thomas Nehrkorn, Joanna Wensell, John Baldwin, Greg McMullin, and Ross Hoffman

[1]Atmospheric and Environmental Research, Inc., Lexington, Massachusetts

1. INTRODUCTION

This paper describes the results of a project supported by the Defense Threat Reduction Agency (DTRA) to provide meteorological datasets that meet specific Atmospheric Transport and Dispersion (ATD) modeling requirements. ATD modeling applications require weather data that is both representative and physically realistic, being composed of basic state parameters that are mutually consistent from a dynamic and thermodynamic standpoint. Additional requirements for this project include for the data to be available at any global location and for time periods that satisfy the particular problem at hand. The modeling system known as the Hazard Prediction and Assessment Capability (HPAC) was developed for use by emergency planners for a variety of ATD applications. It is distributed with two types of climatology data to support long-range ATD planning and incidents for which no other weather information is available. The first type is called probabilistic data, which consists of means and variances for winds, temperature, and relative humidity, and is best used to obtain probabilistic results. These files consist of data for twelve 24-hour periods, each representative of an "average day" for each month of the year. The second type is modal data that describes "typical" conditions, and is best suited for quasi-deterministic forecasts. Modal files consist of 3-dimensional fields of wind, temperature, and moisture derived from the NCEP/NCAR global reanalysis (Kalnay et al. 1996). However, despite its name, the modal data are not representative of modes in the statistical sense; rather, the data have been arbitrarily selected from the 15th and 16th of each month of the year 1990. While the modal data satisfies the "physically consistent" criterion, its arbitrary nature disqualifies this modal data from meeting the "representative" criterion. A truly representative typical day would be the result of a search through a historical record of meteorological conditions to identify values of parameters—or combinations of parameters—that occur with the greatest frequency. During this project, we developed a solution for providing typical day meteorological data for DTRA that meet these criteria by coupling innovative search techniques with long-range historical archives of meteorological data in an effort to recreate a selected historical event identified as representative of typical for any region and season.

Corresponding author address: George D. Modica, AER, Inc., 131 Hartwell Ave., Lexington, MA, 02421-3126; email: gmodica@aer.com

The choice of typical day definition can have a significant effect on downstream applications such as HPAC. Figure 1 is an example showing the result of using two different definitions of typical day to provide meteorological input for an HPAC simulation of a hypothetical Sarin release. One run is based on data from the typical day identified by an early version of the search algorithm for this Pennsylvania location in June (10 June 2000); the other run is based on the typical day modal data provided with the version 4.0 HPAC distribution (15 June 1990). In this case, the differences in the two meteorological conditions translate into dramatically different HPAC solutions.

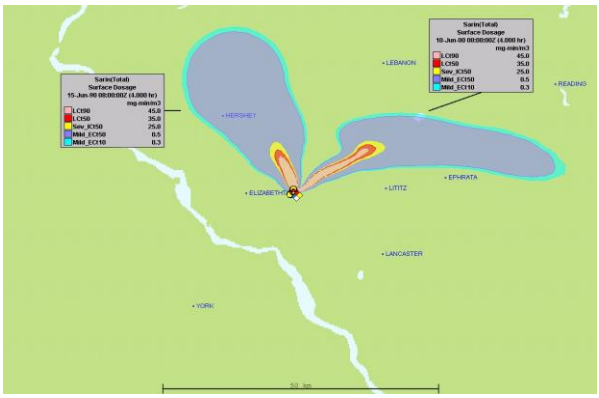

Figure 1. Sarin total surface dosage 4 hours after initial release at 04Z for the dominant mode algorithm (labeled with the date 10 June 2000) and the current HPAC typical day conditions (labeled with the date 15 June 1990).

In Section 2 we describe the Typical Day algorithm and the data used during its development. Section 3 presents the results of tests to demonstrate the integrity of the algorithm to perform consistently under a variety of conditions. A summary and some conclusions are provided in Section 4.

2. ALGORITHM AND DATA

Our solution for providing typical day meteorological data involves the application of tailored search techniques to long-range historical archives for the selection of specific dates that identify typical

conditions for a region and season. The definition of typical is somewhat ambiguous, but is loosely established as a set of conditions, of relevance to the defined problem space, that are most likely to occur at the defined location during the defined season. It should be noted that the more parameters included in an algorithm to define typical, the less likely it will be for any particular combination of those parameters to occur. Therefore there is a tradeoff between selecting events that occur very frequently but are loosely defined (e.g. prevailing wind direction is from the southwest) vs. selecting events based on more precise condition sets of relevance to the ATD problem but that occur less frequently overall. During the initial phase of the project we investigated numerous approaches to this problem and established a prototype system that allowed multiple schemes for identifying and producing "typical" weather to be studied and demonstrated to DTRA.

2.1. Typical Day Baseline Algorithm

Given the desire to focus on the dominant mode of a parameter rather than its mean in establishing "typical," an algorithm based on the successive analysis of univariate histograms was selected as showing the most skill. During early stages, this algorithm was developed, implemented, and tested for two separate locations (in Pennsylvania and Afghanistan), and for four months in different seasons (March, June, September, and December). The algorithm makes use of four quantities derived from the NCEP/NCAR gridded reanalysis data: vector average wind direction, average wind speed, and the minimum and maximum Richardson number (during a 24-hour period) of the lowest layer. These parameters were specifically selected for their relevance to the ATD problem. It should be noted that the same algorithm can be applied with different parameter selections to yield representative typical day meteorological conditions for other problem domains.

The wind quantities are averaged over a 24-hr period to capture one diurnal cycle (four synoptic times in the NCEP/NCAR reanalysis), and over the vertical levels between the surface and approximately 3 km above ground. The vertically averaged wind direction is computed as the direction of the pressure-weighted average wind vector, computed as follows:

$$\overline{V} = \frac{\sum_{k=1}^{K} V_k (p_k^b - p_k^t)}{\sum_{k=1}^{K} (p_k^b - p_k^t)} \qquad (1)$$

where V_k is the wind vector at level k, and

$$p_k^b = \begin{cases} p_k \; if \; k = 1 \\ 0.5(p_{k-1} + p_k) \; if \; k > 1 \end{cases} \quad \text{and}$$

$$p_k^t = \begin{cases} p_k \; if \; k = K \\ 0.5(p_{k+1} + p_k) \; if \; k < K \end{cases}$$

The average wind speed is computed analogously, with V in (1) replaced by $|V|$.

The bulk Richardson number is computed over the lowest available layer, i.e., from the surface to next available level, requiring a minimum layer depth of at least Δpmin = 30 hPa (corresponding to approximately 300 m) as a measure of low-level stability. Because this parameter depends critically on the diurnal cycle, temporal averaging over the entire 24-hour period is inappropriate for this variable. Instead, the maximum and minimum values over the 24-hour period are selected. Because the Richardson number can span a wide range of values and can be strongly skewed, a variable transformation is performed to result in a more regular histogram:

$$Ri' = \ln(\max(Ri, Ro) + 1 - Ro) \qquad (2)$$

where R_o is an adjustable parameter. We used R_o = -5 in most of our algorithm testing and in the examples shown here. The algorithm was applied to the NCEP/NCAR reanalysis data set using 25 years of data from 1979 through 2003, which yields 750 days of data for the June and September analysis (755 for March and December).

The number of bins for the computation of sample histograms is pre-selected based on the sample size. The dominant mode is identified as the bin with maximum density. To avoid selecting spurious isolated maxima from potentially noisy sample histograms, all bins within a given tolerance (we used 20%) of the maximum density value are initially considered as potential dominant modes. The dominant mode is then defined as the bin that contains the largest fraction of the sample population as surrounding bins are added. The sample sub-selection performed for each one-dimensional histogram starts with the bin identified as the dominant mode of the histogram. The sample is then enlarged by adding surrounding bins of the histograms[1], based on one (or both) of the following criteria:

- the fraction of the population represented by the selected histogram bins must exceed a specified minimum value, but should not exceed a specified maximum value, and
- the range of the parameter values of the selected histogram bins must exceed a specified minimum value, but should not exceed a specified maximum value.

In all the results shown here, and the production implementation of the algorithm, we used the first

[1]No bins are added once the end of the data range is encountered, except in the case of wind direction, for which the histograms are circular.

criterion only, requiring a minimum fraction of 1/3, and a maximum fraction of 2/3.

The definition of typical or representative is ultimately based on a measure of distance between the meteorological state for the day in question and the central values of the dominant mode. For a multivariate analysis such as ours, the choice of scaling and distance metrics thus becomes an integral part of the algorithm. For the histogram analysis based on the pre-selected variables, distances from the dominant mode were based on a normalized Cartesian distance *d*

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - m_i)^2}{s_i^2}} \qquad (3)$$

where the summation extends over all variables (e.g., for wind direction, minimum R_i', maximum R_i', and wind speed), the x_i are the values of these quantities for the day in question, the m_i are the corresponding dominant mode central values, and the s_i are the scales for these variables. We used the inter-quartile range over the entire sample as the scale for each variable, but required scales at least as large as the histogram resolution in the analysis presented in the production implementation of the algorithm.

2.2. Data

We extracted pressure-level data from a 25-year (1979-2003) collection of NCEP/NCAR at the locations illustrated in Figure 2. These locations were selected because they include tropical, sub-tropical, mid-latitude, continental, and maritime regions, and locations at high elevation as well as near sea level. By selecting locations that cover a wide range of meteorological conditions, we were able to "stress" the code to reveal vulnerabilities in the baseline algorithm.

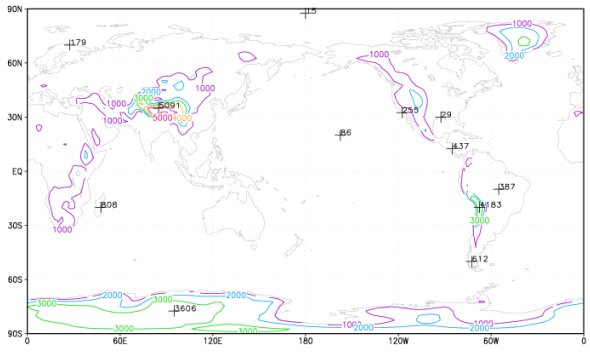

Figure 2. Locations selected to test the Typical Day algorithm before implementation into the production architecture. Locations marked with a '+' and are followed by the site elevation (m; contour interval is 1000 m).

Data was collected for the months of March, June, September, and December in order to provide representative conditions for each season. Additional data was collected for the months of January May, July, and November to enable us to test two different contiguous 3-month periods (i.e., November-December-January and May-June-July).

3. RESULTS

3.1. Baseline Algorithm Evaluation and Results

The baseline histogram algorithm outlined in Section 2.1 was run for the locations shown in Figure 2, and several quantitative measures of algorithm performance were documented. Graphical histograms provided an economical way to verify whether or not the algorithm behaved as expected. Only data from the Brazil location for the month of September is presented in the following discussion, although we note this example was representative of the other locations and seasons evaluated.

The histogram sequence for Brazil in September is shown in Figure 3. The algorithm was provided with input data from an initial sample of 750 September days. The upper left plot is the graphical result for the averaged wind direction histogram test. The mode (the most frequently occurring value) is indicated with a red line and is also labeled above the plot (90). The number that follows the labeled mode after the colon (287) represents the number of days, or sub-selections, from the original 750 that were passed on to the next histogram test. The green lines indicate the range of wind direction encompassed by the sub-selections. The next two histogram tests that follow are for the transformed Richardson number (2), with 122 and 76 sub-selections, respectively. In the lower-right plot, 43 days remain after the averaged wind speed histogram test.

While the metric *d* in (3) provides an objective measure of the baseline algorithm performance, it is also reassuring to observe the result in terms of common meteorological variables. Toward that end, vertical profiles of wind direction, wind speed, and temperature at the four times available in the NCEP/NCAR data set, are shown in Figure 4 for the highest-ranked typical day identified by the baseline algorithm (25 September 1985). For comparison, vertical profiles are also shown for the dataset currently serving as "Typical Day" provided with HPAC (15 September 1990). For this particular case, the two typical days reveal differences that could be important for ATD: for the dominant mode algorithm, representative conditions are light winds from the east and a minimal diurnal cycle of low-level static stability, whereas the HPAC typical day has weak southerly winds and indicates a strong nocturnal inversion.

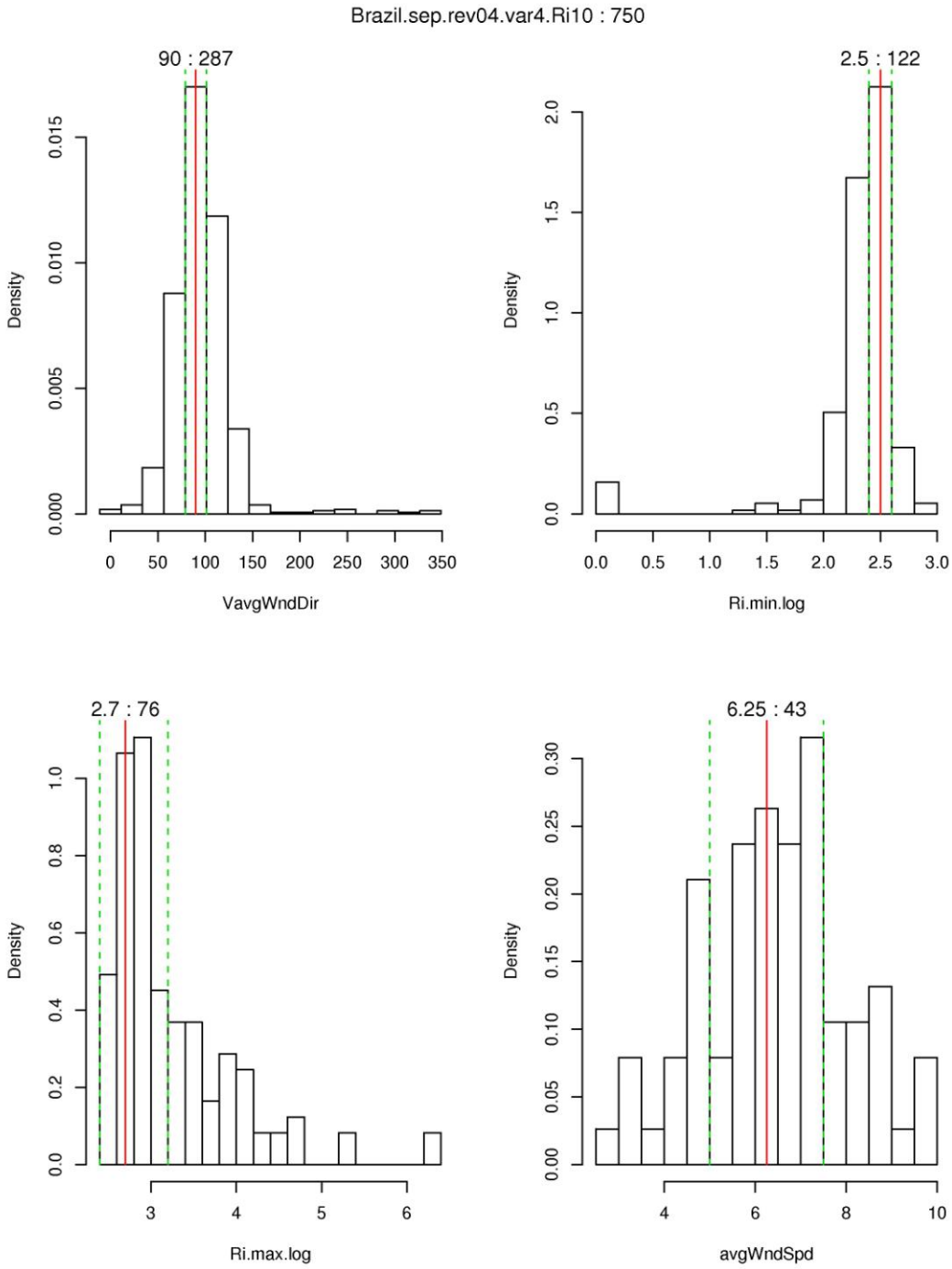

Figure 3. Dominant mode histograms for Brazil for all September months (25) during the period 1979-2003. Variables used in the baseline algorithm are averaged daily wind direction (upper left), the 24-hour minimum in the Richardson number (upper right), the 24- hour maximum in the Richardson number (lower left), and the averaged daily wind speed (lower right). The red line indicates the mode selected by the algorithm, which is also labeled at the top of each histogram along with the number of sub-selections. The green dashed line indicates the width of the sub-selection.

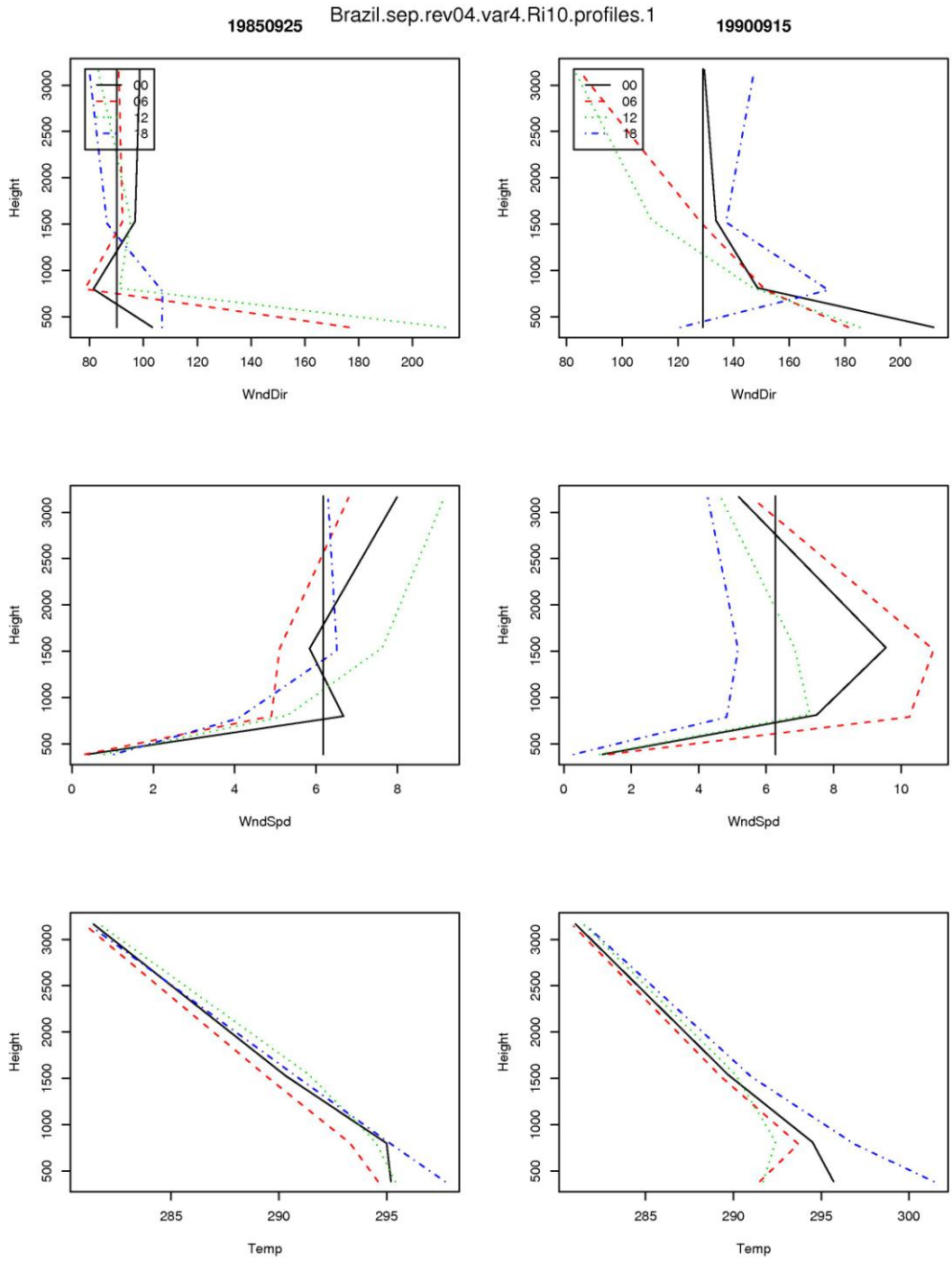

Figure 4. Vertical profiles of wind direction (degrees; top), wind speed (ms-1; middle), and temperature (K; bottom) for the typical day identified by the dominant mode algorithm (25 September 1985; left), and the HPAC typical day (15 September 1990; right), for the Brazil location. For wind direction and speed, vertical lines indicate the vertically and temporally averaged value.

The baseline typical day algorithm operates on a 31-day sample selected from a desired month of the year, leading to the result that there will be 12 typical days defined for a given location. In this regard the sample of days is effectively "centered" at mid-month (about the 15th of the month). While there were some differences in typical day results when historical samples are defined with respect to time periods other than monthly samples (not shown), the typical meteorological conditions as defined by the centroid values of the selected variables are robust. The resulting typical day selections were also found to be relatively robust, although some cases were found in which the selection of days was more sensitive.

In the final step, the algorithm sorts the surviving typical day selections—in this case, 43—based on each day's normalized distance from the computed centroid [*d* in (3)]. To gain an appreciation of the relative sizes of the values of *d*, the data is presented graphically in Figure 5. Notice the difference between *d* for each of the top-10 days selected by the baseline algorithm and that for the original typical day definition provided with the HPAC v4.01 application (labeled "19900915" in the figure). This difference is an indication that the days selected by the histogram algorithm are much closer to the mode than the arbitrary 15th day of the month, and therefore better represent the most frequent, or typical meteorological conditions at this desired location and time.

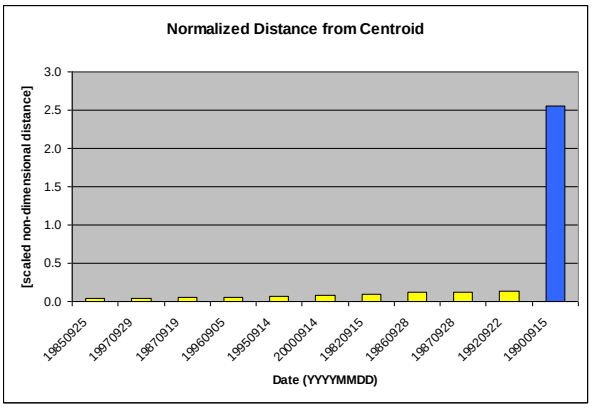

Figure 5. Illustration of the normalized distance from the centroid to each of the top-10 typical September days selected by the histogram algorithm (yellow), and from the centroid to the current HPAC definition of Typical Day (September 15, 1990; blue) for the Brazil location during the 1979-2003 period.

3.2. Compact Typical Day Datasets

During execution of the baseline algorithm on the full historical dataset, the centroid in the four-dimensional space of the spatio-temporally averaged variables is identified by the central values of the dominant mode histogram bins. In addition, the subsample of days contained within these dominant modes is ordered by normalized distance from this centroid, and the closest one is identified as the typical day from the historical sample. However, since by definition the typical day is characterized by frequently occurring atmospheric conditions, it is to be expected that a similarly typical day can be found in the corresponding month of other years, if not every year. For this reason, a typical day dataset can be made to include only a subset of years in the gridded dataset. To determine the minimum acceptable number of years, we examined the distribution of the normalized distance from the typical day centroid for different combinations of 1, 2, and multi-year datasets.

We considered all possible subsamples of consecutive 1, 2, or multi-year datasets (we restricted this analysis to consecutive years because this allows for longer-period simulations when near the beginning or end of the year). For each location and month, we determined the number of typical days (i.e. days that fall within the required proximity of the centroid) in all single- and multi-year subsets. The number of instances in which no typical days were found within a subset was tallied separately for 1, 2, and multi-year datasets. We first considered only days within the same calendar month as in the historical sample as candidate typical days. Using this criterion, we found that a minimum of 4 years are needed to ensure that a typical day can be found for each month, location, and consecutive 4-year dataset. When an expanded matching period of 61 days (+/- 30 days from the 16th of the month) was considered, we observed that a three-year dataset was sufficient to ensure a typical day can be found for all instances of location, month, and dataset.

Assuming the data we selected for this exercise are representative, the expected probability of encountering no instances of a typical day for a dataset of given length is shown in Figure 6. For the 31-day matching period used in the baseline algorithm, the results indicate that a dataset of 4-year length will likely always yield a typical day; the probability of encountering a zero typical day instance is still <1% for the 3-year dataset, but increases rapidly to 0.9% and 2.0% for the 2- and 1-year datasets, respectively. The 61-day matching period provides even better odds of providing a typical day, with the 3-year dataset more likely to yield a typical day than its 31-day counterpart.

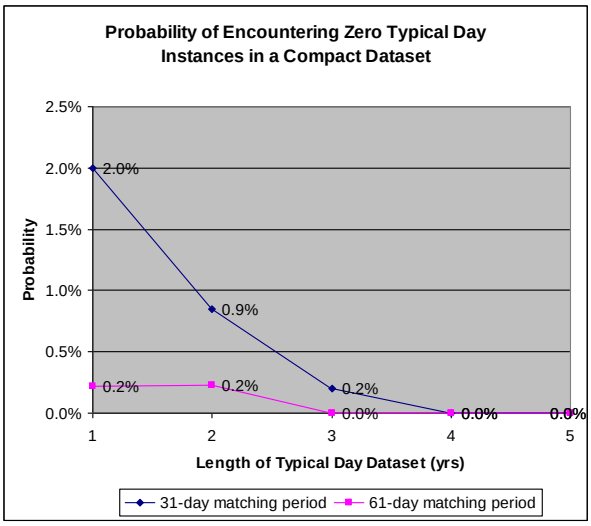

Figure 6. Graphical representation of the probability of encountering zero typical day instances as a function of typical day dataset length for the 31-day (blue) and 61-day (magenta) matching periods.

One last examination of the applicability of the compact TD dataset involved all the gridpoints in the 2.5° × 2.5° NCEP/NCAR Global Reanalysis. It focused on the result of restricting the filtered typical day to the three-year period 1989-1991, rather than choosing the top-ranked date from the entire 25-year sample 1979-2003. We computed the rank of the filtered typical day (for each location and month) within the corresponding 25-year sample. The rank was determined for each day in the 25-year sample list by ordering the dates in ascending order of normalized distance from the centroid. For this ranking, all dates within the final selection were considered first, followed by all dates outside the final selection. A corresponding percentile was computed by multiplying the rank by 100, divided by the size of the whole sample.

The percentiles were computed for all grid locations and all months (of the total 126,144 typical days, 9395 (7.4%) of the metrics were unavailable because of processing problems). Overall, for the vast majority of cases, the filtered typical day is very near the top-ranked dates for the entire sample: for more than 99% of the analyzed typical days, the filtered date is within the top-ranked 5% of the entire sample. The median percentile is 0.77%, with a slightly higher mean (1.17%) due to isolated cases with higher percentiles. In the worst case, the filtered typical day was still within the top 33% of the entire sample (see the attached histogram plot for full details). Stratification of these statistics by month and geographical region (we considered 30-degree latitude bands in the North and South Hemisphere) revealed no systematic patterns of particularly low or high percentile values. Maximum percentile values show the greatest variation, but these represent an insignificant fraction of the sample. For each month and region, there was at least one point in which the top-ranked date fell within the filtered years. Individual mean and median values are all quite close to those for the grand total of all locations and months.

4. SUMMARY AND CONCLUSIONS

We developed a Typical Day algorithm and tested an implementation on NCEP/NCAR Global Reanalysis data. During tests of the dominant mode baseline algorithm, we observed considerable skill in the algorithm's ability to identify representative typical day weather conditions of relevance to the ATD problem. Furthermore, the algorithm is particularly well suited to the job of producing a global HPAC typical day dataset because it can be executed in a fully automated environment without a human-in-the-loop, and performs equally well at for any location and season.

Acknowledgements: This work was supported by the Defense Threat Reduction Agency under contract HDTRA1-07-C-012. The NCEP Reanalysis data were obtained from the NOAA-CIRES Climate Diagnostics Center, Boulder, CO from their web site at http://www.cdc.noaa.gov.

REFERENCES

HPAC Users Guide, 2003, SAIC.

Kalnay, E. and others, 1996: The NCEP/NCAR 40-Year Reanalysis Project, *Bull. Amer. Met. Soc.*, **77**, 437-471.