Climatic data analysis for input to ShipIR

David A. Vaitekunas[a,1], Yoonsik Kim[b,2]
[a]W.R. Davis Engineering Limited, 1260 Old Innes Road, Ottawa, Ontario, Canada K1B 3V3
[b] Korea Institute of Ocean Science and Technology, Daejeon, Republic of Korea

Proc. SPIE 8706, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXIV, 87060Q (June 5, 2013); doi:10.1117/12.2016498)

ABSTRACT

A key input to any thermal infrared signature model is the environment, more specifically the model inputs specific to the thermal infrared background model. This paper describes a new method of analysing the climatic data for input to ShipIR. Historical hourly data from a stationary marine buoy are used to select a small number of data points (N=100) to adequately cover the range of statistics (CDF, PDF) displayed by the original data set (S=46,072). The method uses a coarse bin (1/3) to subdivide the variable space (3^5=243 bins), and a single-point ranking system to select individual points so that uniform coverage (1/N = 0.01) is obtained for each variable. The selected data points are used in Vaitekunas and Kim (2013) to demonstrate how the new methodology is used to provide a more rigorous and comprehensive analysis of platform IR susceptibility based on the statistics of IR detection.

Keywords: platform signature, environments, effects of climate, climate statistics, selection algorithm

1. INTRODUCTION

ShipIR/NTCS is a comprehensive software engineering tool for predicting the thermal infrared (IR) signature and IR susceptibility of naval warships. The ShipIR component consists of several sub-models, including the MODTRAN4 infrared sky radiance and atmosphere propagation model, a proprietary sea reflectance model combining the methods of Mermelstein (1994) with the results from Shaw and Churnside (1997) and Ross and Dion (2007). The platform model is created from a 3D surface geometry that forms the basis of both a radiative heat transfer and in-band surface radiance model comprised of diffuse and specular multi-bounce reflections. An exhaust plume trajectory and IR emission model predicts the infrared signature of diesel engine and gas turbine exhaust systems. Internal heat sources are modelled via user-defined thermal boundary conditions, simulating a complex thermal network of specified temperatures (controlled spaces), forced and natural convection, heat-flux, and heat conduction. Validation of the ShipIR model has been the topic of several research papers (Vaitekunas and Fraedrich 1999, Fraedrich et al. 2003, Vaitekunas 2005).

The main purpose of climatic data in an infrared analysis is to model the effect on both ship signature and infrared sensor detection. In previous studies, a set of four backgrounds were typically used, denoting the worst possible (wp) and best possible (bp) day (d) and night (n) operating scenarios. Actual conditions were obtained using an analysis of the baseline ship (skin-only) with monthly averages taken from the US Navy Marine Climatic Atlas of the World (USNMCAW) for a specific area of interest (e.g., Eastern Sea). Realising the previous analysis did not consider the variance in the monthly statistics, attempts were made to analyse the standard deviations stored in the USNMCAW (Vaitekunas 2010). These previous results were incomplete for two reasons:

- the method did not analyse or account for the inherent correlation between each climatic variable,

[1] dvaitekunas@davis-eng.com; http://www.davis-eng.com; phone: +1 613 748 5500; fax: +1 613 748 3972

[2] yoonsik@kiost.ac; http://www.kiost.ac; phone: +82 42 866 3454; fax: +82 42 866 3449

- pre-selection of values for each variable to produce a minimum, average, or maximum IR signature and IR susceptibility failed to produce the desired result (i.e., average conditions showed a higher signature and IR susceptibility than the worst or highest signature condition).

This paper will describe how a more rigorous method was developed to select the climatic data using existing historical data, in this case a stationary marine buoy operated by the Korea Meteorological Administration (KMA) located in the Eastern Sea (37.5°N, 130.0°E). The objective of the new method is to take a small sample (N=100) from a much larger historical data set (S=46,072) such that each individual variable is uniformly distributed in CDF (1/N = 0.01) over its range in the original data while adequately sampling any non-uniformity in the distribution PDF. The underlying objective is to provide an adequate range of conditions under which a new or existing ship will operate, to capture the true expected value, E(x), and the 5th, 50th and 95th percentiles in IR susceptibility.

2. METHODS AND RESULTS

Figure 1 contains 4 of the 10 possible 2D histograms for the five KMA climatic data variables: air (Ta) and sea (Ts) temperature, relative humidity (RH), wind speed (Ws) and direction (Wd). The warmer colours indicate a higher frequency or probability of occurrence of the data pair. Higher frequencies concentrated along a curve, y = f(x), indicate a high degree of correlation between the variables (e.g., air and sea temperature). These results are significant since they constitute the first evidence of correlation between the data. To further quantify the correlation or dependency, Table 1 presents the correlation matrix as computed by the Analysis ToolPak in Microsoft Excel. These results show a varied amount

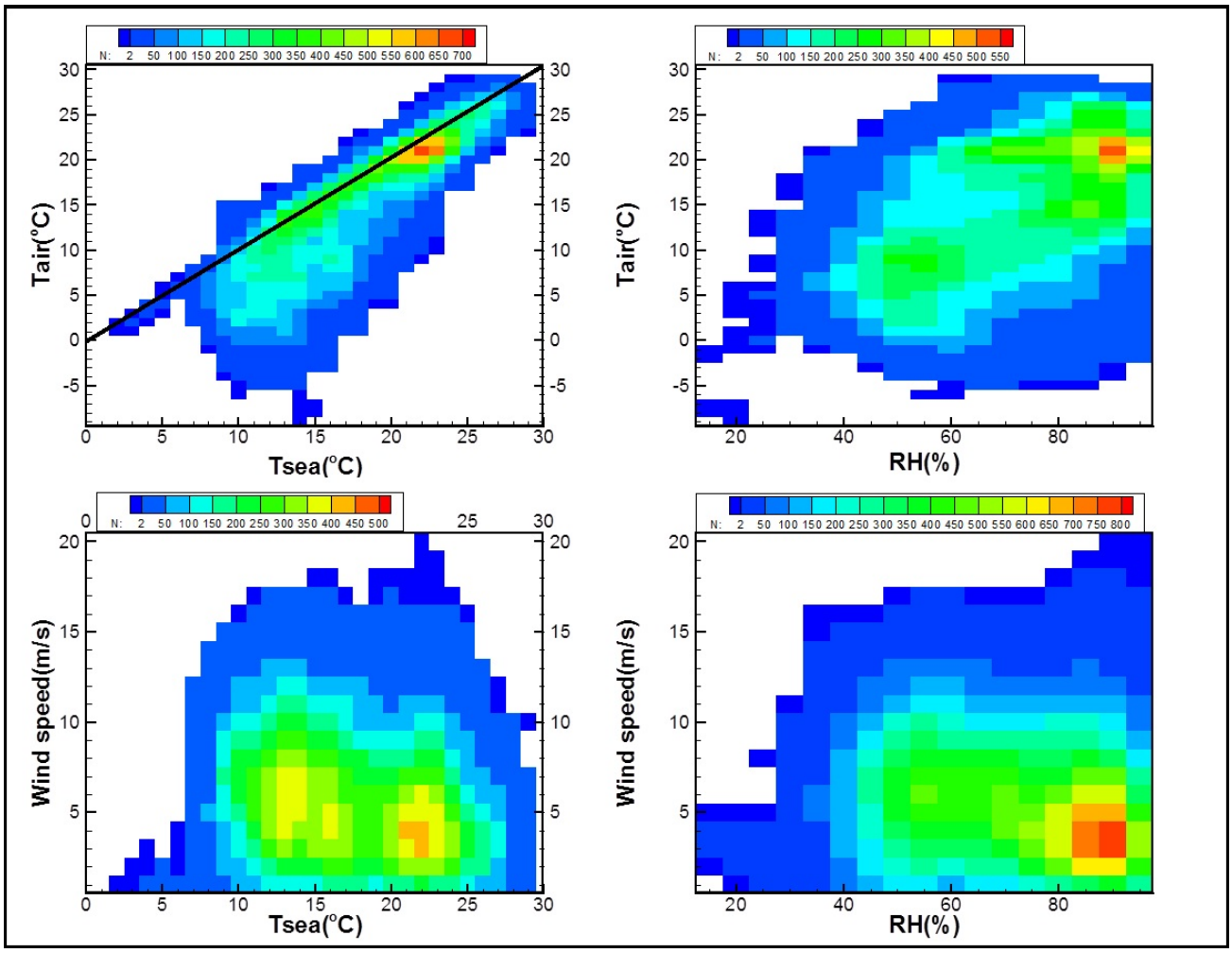

Figure 1: 2D histograms for various pairs of data in the KMA marine buoy data.

of correlation between the variables, ranging from 0.12 (low) to 0.86 (high). The sections to follow will describe the three step process of coarse binning, single-point ranking, and point selection used to obtain the N data points used in our IR susceptibility analysis.

Table 1: correlation matrix for entire KMA data set.

	Ta(oC)	*Ts(oC)*	*RH(%)*	*Ws(m/s)*	*Wd(true)*
Ta(oC)	1				
Ts(oC)	0.857595	1			
RH(%)	0.530476	0.322869	1		
Ws(m/s)	-0.32284	-0.20673	-0.14502	1	
Wd(true)	-0.20563	-0.16321	-0.2424	0.124879	1

2.1 Coarse Binning

The large data set (S=46,072) is divided into large clusters containing 1/3 of each variable cumulative distribution (CDF), which translates into $3^5 = 243$ potential bins for the five variables (Ta, Ts, RH, Ws, Wd). The purpose of coarse binning is to quickly isolate manageable collections of data to choose from with some degree of certainty about their correlated probability of occurrence. The cluster size (1/3) was chosen simply based on the number of points ($N = 100 \leq 3^5$). The coarse bins cover a large enough fraction of the total volume that their correlated probability can be estimated from n/S, where n is the number of data in the bin and S is the size of the large data set (46,072). Figures 2 and 3 show the individual probabilities versus an arbitrary bin index, and their size distribution from smallest to largest, respectively. These results show that the frequency within each of these coarse bins varies by two orders of magnitude, and 45 of the 243 bins (18.5%) are empty (n=0).

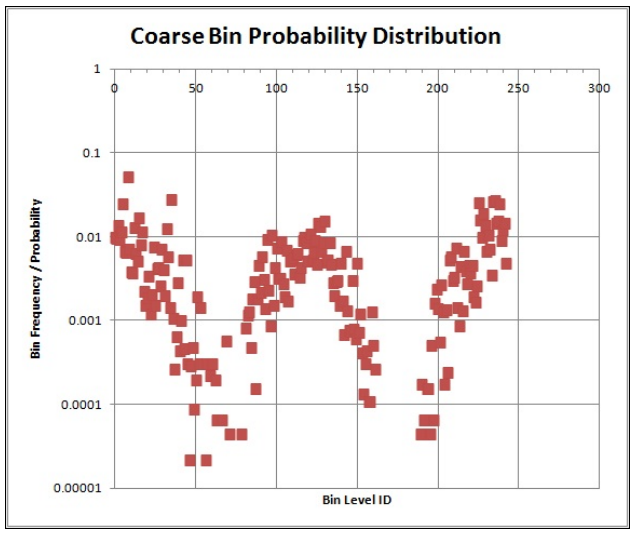

Figure 2: coarse bin probability distribution.

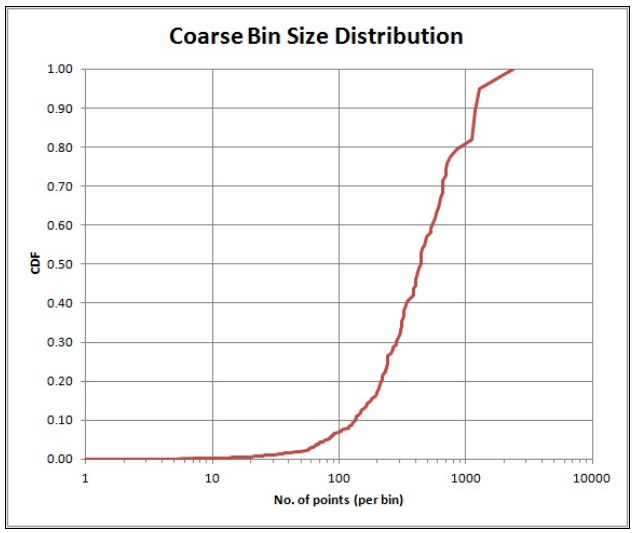

Figure 3: coarse bin size distribution.

2.2 Single-Point Ranking

Contrary to coarse binning where large volumes of data are coalesced to compute a single probability (frequency), the purpose of single-point ranking is to score individual points based on their suitability to fill a void in the data requirement. Since the goal is to maximize coverage of individual variable CDF, the first obvious metric is the number of unused bins in each variable CDF (1/N). Given the two orders of magnitude difference between the low-probability and high-probability regions of the variable space (Figure 2), the low-probability points should be selected first, employing the principle of *grab them while you can*. A method is therefore required to calculate the probability of occurrence for individual data points based on their individual CDF.

The simplest way to calculate a joint probability in *m* dimension space is to take the product of individual (uncorrelated) probabilities for each variable. Since all 5 variables are correlated to different degrees (see Table 1), a method is needed to decorrelate the variable space. Principle Component Analysis or PCA, a method commonly used to identify the most meaningful basis to re-express a data set (Shlens 2009), is well suited to decorrelate the data. The method uses a linear combination of the original data space X={Ta,Ts,RH,Ws,Wd} to define a new variable space $Y=\{Y_1,Y_2,Y_3,Y_4,Y_5\}$ such that the covariance is a minimum and each successive dimension in Y is rank-ordered by its variance. In our application, the order

```
load pca-in.txt;
[N,M] = size(pca_in);
mn = mean(pca_in);
X = pca_in - repmat(mn,N,1);
Cov = 1 / (N - 1 ) * X' * X;
[PC,V] = eig( Cov );
V = diag(V);
[junk,rindices] = sort( -1*V );
V = V(rindices);
PC = PC(:,rindices);
Y = X * PC;
save -ascii pca-out.txt Y;
save -ascii pca-mat.txt PC;
save -ascii pca-cov.txt Cov;
save -ascii pca-var.txt V;
```

Figure 4: octave script for PCA

Table 2: PC transformation matrix.

	Ta(oC)	Ts(oC)	RH(%)	Ws(m/s)	Wd(true)
Y1	0.000412	-0.006106	0.053498	0.140470	-0.988620
Y2	0.001635	-0.019222	0.180679	0.972148	0.148026
Y3	0.004852	-0.024269	0.981684	-0.187006	0.026703
Y4	-0.012612	0.999421	0.027700	0.015022	-0.002544
Y5	-0.999907	-0.012758	0.004732	0.000550	-0.000004

Table 3: correlation matrix for the Y values.

	Y1	*Y2*	*Y3*	*Y4*	*Y5*
Y1	1				
Y2	-4.8E-12	1			
Y3	-1.2E-11	-1.1E-11	1		
Y4	-3.7E-12	5.48E-13	-1.4E-11	1	
Y5	-1.9E-11	6.27E-12	-3.7E-12	1.54E-12	1

of the Y variables is not important since only the probability is needed for point ranking[3]. Figure 4 contains the octave script used to perform the PCA. The resulting linear transformation (PC) matrix and correlation on Y are shown in Tables 2 and 3, respectively. The Y results are clearly uncorrelated. Using the uncorrelated Y, the joint probability is calculated for each data point using the product of individual (uncorrelated) probabilities in Y. The following discrete equation is used to obtain the probability in each *k* dimension of Y using the individual CDF from $Y_{k,min}$ to $Y_{k,max}$:

$$P_i(Y_k) = \frac{CDF_{k,i+1} - CDF_{k,i-1}}{Y_{k,i+1} - Y_{k,i-1}} \quad k = 1,5 \quad i = 1,S \tag{1}$$

The results are then mapped to the original data point (unsorted) using its point ID, and multiplied by each k dimension:

$$P_i = P_i(Y_1) \cdot P_i(Y_2) \cdot P_i(Y_3) \ldots \cdot P_i(Y_5) \quad i = 1,S \tag{2}$$

and normalized to obtain a joint discrete probability value for each data point:

$$P_{tot,i} = \frac{P_i}{P_{tot}} \quad P_{tot} = \sum_i^S P_i, \quad i = 1,S \tag{3}$$

The following equation is now used to rank each point:

$$R_i = \frac{m}{P_{tot,i}} \tag{4}$$

m is the number of slots available in the 1/N CDF of (Ta, Ts, RH, Ws, Wd). All the variables start out with a value of *m* greater than or equal to 5. Some points will have a larger value of *m* if more than 1/N of the data share the same value. This occurs when the measurement resolution is insufficient to discriminate to within the desired resolution in CDF. One example is relative humidity which ranges from 12% to 100% with a resolution of 1% (88 bins). Since the available slots will diminish with each point selection, the above ranking needs to be recalculated at every step in the point selection process.

2.3 Point Selection Algorithm

A single data point (per coarse bin) is selected such that the resultant data set (N=100) has no repeats in the CDF (1/N = 1/100) for each climatic input variable. As described, each point is ranked based on its unlikelihood to reoccur (1/Ptot) and the degree to which it fills an available slot in the CDF. The first attempt to implement the algorithm used an excel spreadsheet to sort the remaining unused bins (from least probable to most probable) and rank all the points within the next available bin. The coarse bins and data points with the least probability of occurrence were selected first since they are unlikely to fill any unused slots in the CDF later in the process. Since the first attempt involved manual intervention at every

[3] Originally we tried to use the PCA to select the data points by randomly generating a combination of Y values and transforming them back into X. The results were invalid because the independent selection of Y does not provide any control over where the points end up in the original variable space (a by-product of uncorrelating the data).

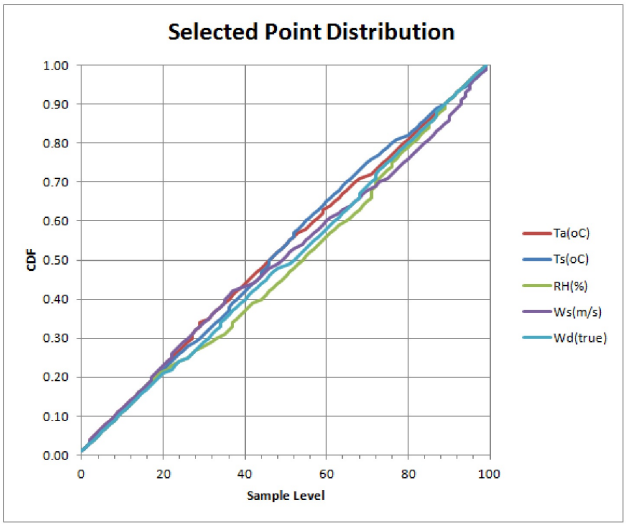

Figure 5: resultant variable CDF from manual point selection.

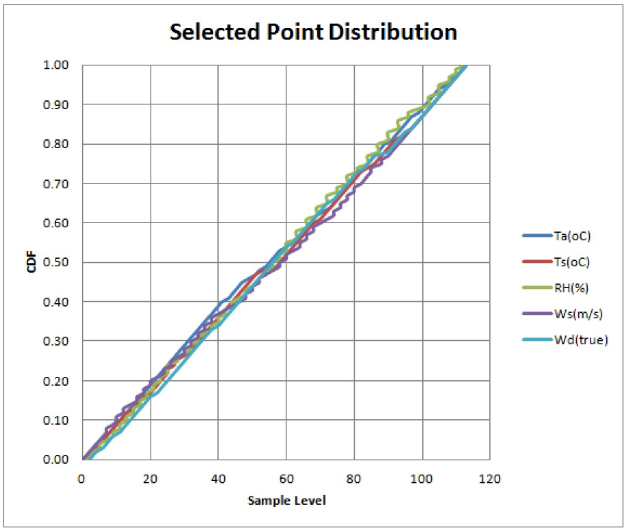

Figure 6: resultant variable CDF from automated point selection.

step in the process, no allowance was made for reiteration. As a result, some repeats in discrete CDF were allowed but their occurrence was carefully controlled (manually) to make sure repeats only occurred once for the same CDF. The algorithm has since been implemented as a computer program, and further enhancements were added to avoid any repeats in CDF. In addition to re-ranking the data points in the bins already processed, and swapping points with a higher ranking (later in the process), the method was found to be less biased towards the end if the coarse bins are randomly chosen from both the low probability (CDF<0.5) and high probability (CDF>0.5) regions, in alternate succession. Repeats were avoided by simply increasing the initial size of the selection bin (N) until the closest number of points to the desired value (100) is obtained. In the KMA data set, an initial value of N=114 produced 101 data points with no repeats. Figures 5 and 6 show the resultant CDF values per sample level (1-100) for the manual and automated point selection algorithms, respectively. A perfect selection would be five (5) overlapping straight lines at 45°with one increment in CDF per sample level. Repeats in one variable cause a vertical step change in CDF for the same sample level, while missing values cause a horizontal step in value (sample level) for the same CDF. The closer these curves are to the ideal (45°) line, the better the selection algorithm performs on the data set. The benefits obtained through refinements in the automated algorithm are clearly obvious.

2.4 Resulting Statistics

The CDF and PDF obtained from the entire data set, and interpolated using the N=100 data points, are shown in Figures 7 through 18. These results provide the first evidence that all the distributions are non-Gaussian and that each distribution has unique features likely to change from one geographic region to another (since they are rooted in the data set). Air and sea temperature and RH have two distinct peaks in probability. Whereas sea temperature and wind speed distributions are skewed towards the low end of their scales, air temperature and RH are skewed towards the high end of their scales. The Wind speeds follow a typical Weibull (Gamma) distribution where a majority of the data is recorded at moderate levels while high winds do occur but less frequently. Winds in the Eastern Sea are shown to have two prevalent directions (South and Northwest). Although not selected directly, the statistics for the air-sea temperature difference (ASTD) are shown in Figure 18, and although the curves are not as smooth as the other Figures (a different calculation method was used to calculate the CDF and PDF[4]), the results do illustrate two more features about the analysis: a) the distribution of the dependent variables (e.g., ASTD) are equally covered by the point selection process, b) the peak in ASTD is centred about -1°C – a fact to consider when analysing the effectiveness of hull cooling using sea water spray (Vaitekunas and Kim, 2013).

[4] The automated computer analysis includes a special calculation of CDF and PDF which uses a minimum no. of samples (50) to calculate the average CDF and avoids any bias introduced by the minimum measurement resolution.

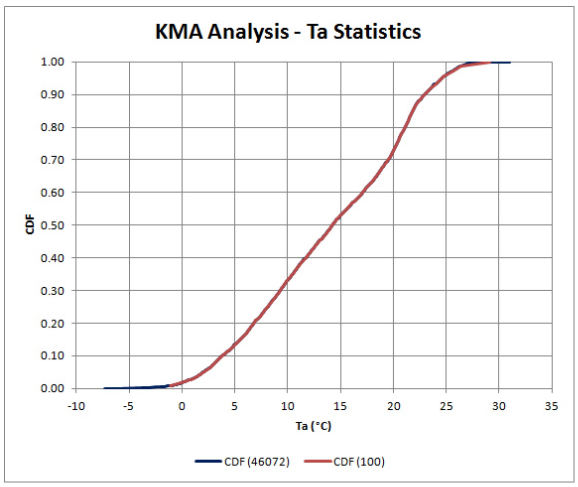

Figure 7: sampled versus data set CDF for Ta.

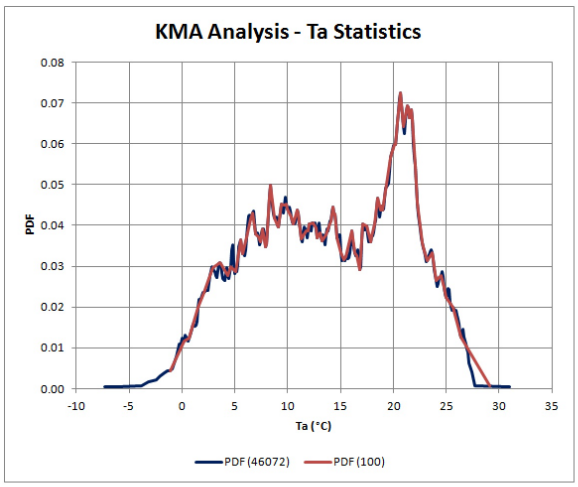

Figure 8: sampled versus data set PDF for Ta.

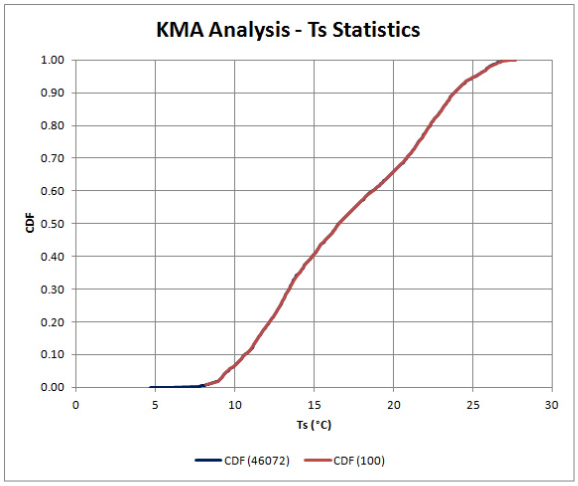

Figure 9: sampled versus data set CDF for Ts.

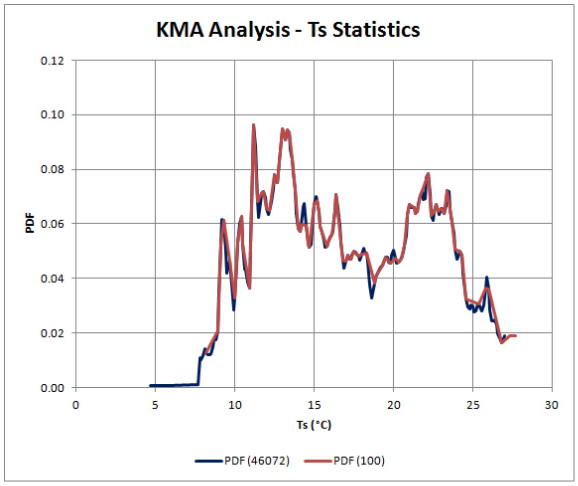

Figure 10: sampled versus data set PDF for Ts.

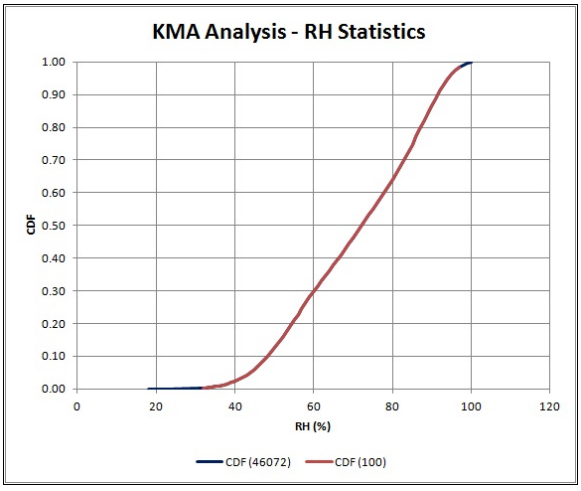

Figure 11: sampled versus data set CDF for RH.

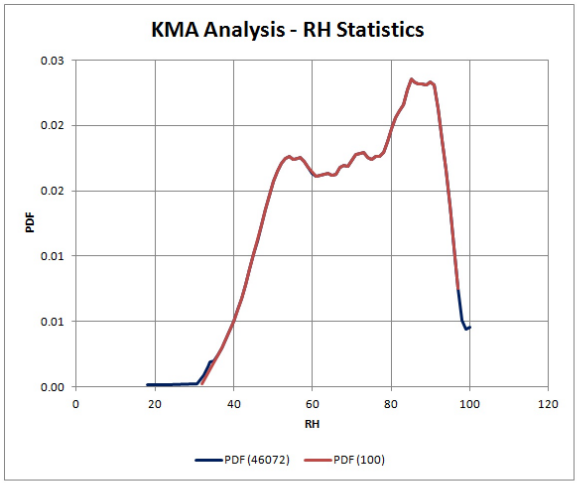

Figure 12: sampled versus data set PDF for RH.

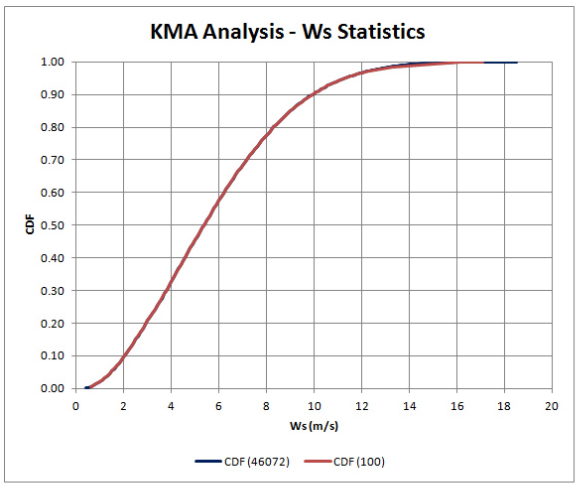

Figure 13: sampled versus data set CDF for Ws.

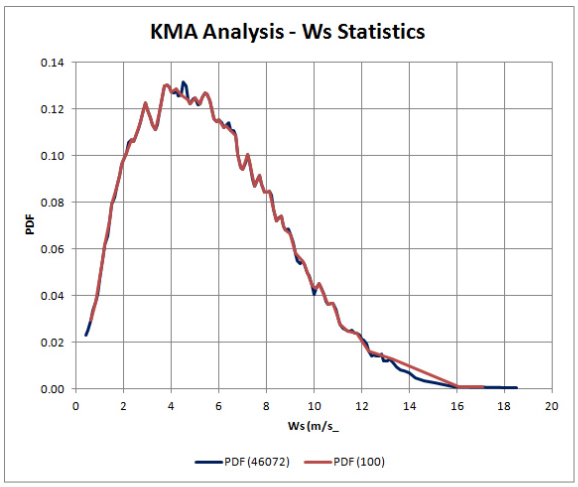

Figure 14: sampled versus data set PDF for Ws.

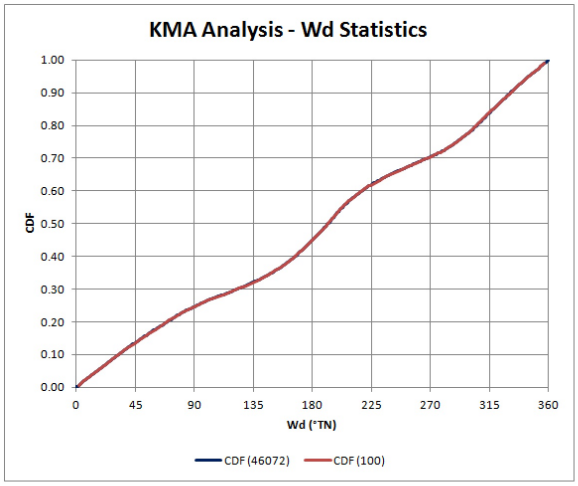

Figure 15: sampled versus data set CDF for Wd.

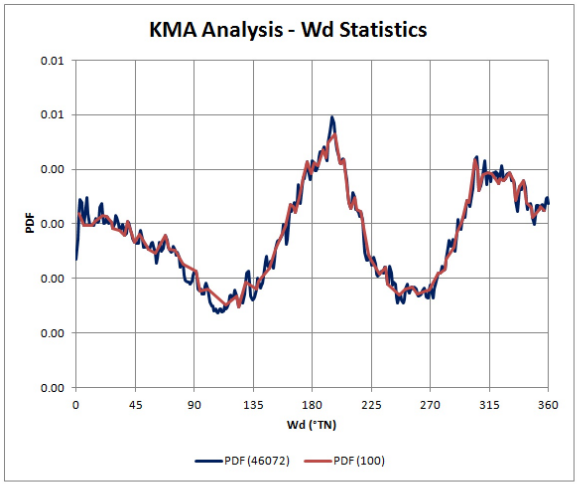

Figure 16: sampled versus data set PDF for Wd.

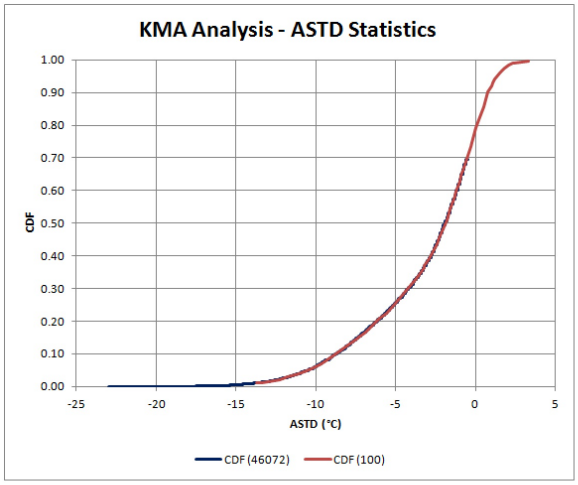

Figure 17: sampled versus data set CDF for ASTD.

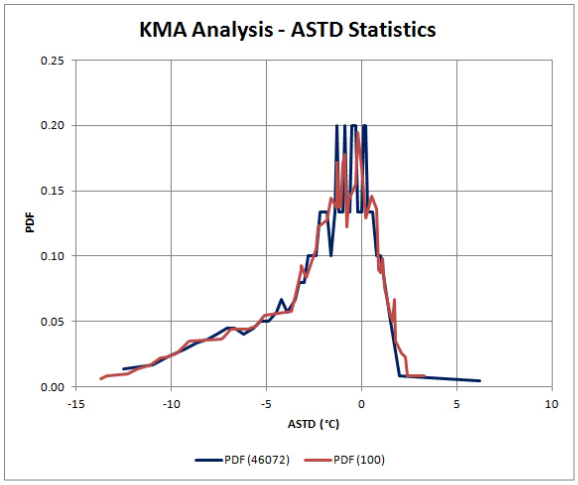

Figure 18: sampled versus data set PDF for ASTD.

3. SUMMARY AND CONCLUSION

A new methodology of analysing historical recordings of climatic data has been presented. The method uses data from a stationary marine buoy to select a small number of data points (N=100) to adequately cover the range of statistics (CDF, PDF) displayed by the original data set (S=46,072). Both a manual and automated point selection algorithm were described which utilize coarse binning (243 bins) and single-point ranking to identify the most likely candidates to fulfill the data requirement for a small representative data set. The single-point ranking system uses Principle Component Analysis (PCA) to decorrelate the input variables and compute a joint probability of occurrence for each data point. Special care was taken to handle situations where the CDF bin resolution was higher than the measurement resolution (e.g., RH), and further remove any potential bias from the procedure as the number of available bins diminish. The resultant data points are used by Vaitekunas and Kim (2013) to provide a more rigorous and comprehensive analysis of platform IR susceptibility based on the statistics of IR detection.

4. FUTURE WORK

Some of the existing meteorological ground stations (closest to the marine buoy) include a global solar radiation sensor (pyranometer) and thermal radiation sensor (pyrgeometer). The MODTRAN atmosphere model could be used to test for clouds, and determine a suitable cloud altitude and extinction value for input to ShipIR. The methods used to filter the historical data for bad readings and dead bands (i.e., wind speed and direction) and unbias the resultant data (after the removal of bad data) will be discussed in another paper, which will also compare the statistics from different climatic regions (e.g., North Atlantic, Persian Gulf, North Sea).

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge the contributions made by the following persons. Dominic Muzar from the University of Ottawa, who implemented the point selection algorithm into software during his work term with W.R. Davis Engineering, in partial fulfilment of his Undergraduate Coop program, and Tom Davis and Dr. Srini Ramaswamy from W.R. Davis Engineering for their useful discussions on the topic, in particular the pitfalls of using the PCA inversion to select the data points.

6. REFERENCES

1. Cox, C. and Munk, W., "Measurement of the Roughness of the Sea Surface from Photographs of the Sun's Glitter," *J. Opt. Society Am.* 44, 838-850 (1954).
2. Fraedrich, D., S., Stark, E., Heen, L.,T., and Miller, C., "ShipIR model validation using NATO SIMVEX experiment results," *Proc. SPIE* 5075, Targets and Backgrounds IX: Characterization and Representation, 49-59 (2003).
3. Mermelstein, M., D., Shettle, E., P., Takken, E., H. and Priest, R., G., "Infrared radiance and solar glint at the ocean-sky horizon," *Appl. Opt.* 33 (25), 6022-6034 (1994).
4. Ross, V. and Dion, D., "Sea surface slope statistics derived from sun glint radiance measurements and their apparent dependence on sensor elevation," J. Geophys. Res., 112, C09015, doi:10.1029/2007JC004137 (2007).
5. Shaw, J.,A. and Churnsize, J.,H., "Scanning-laser glint measurements of sea-surface slope statistics," *Appl. Opt.* 36 (18):4202-4213 (1997).
6. Shlens J., "A Tutorial on Principal Component Analysis," http://www.snl.salk.edu/~shlens/pca.pdf (2009).
7. Vaitekunas, D., A. and Fraedrich, D., S., "Validation of the NATO-standard ship signature model (SHIPIR)," *Proc. SPIE* 3699, Targets and Backgrounds: Characterization and Representation V, 103-113 (1999).
8. Vaitekunas, D., A., "Validation of ShipIR (v3.2): methods and results," 1st International Workshop for IR Target and Background Modelling 27–30 June Ettlingen Germany (2005).
9. Vaitekunas, D., A., "IR susceptibility of naval ships using ShipIR/NTCS," *Proc. SPIE 7662*, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXI, 76620V (April 22, 2010); doi:10.1117/12.852131.
10. Vaitekunas, D., A., and Kim, Y., "IR signature management for the modern navy," Proc. SPIE 8706, Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXIV, (2013).