Developing accurate Multivariate Linear Regression (MLR) models for meteorologically influenced response variables with focus on extreme events

W. E. Frick* and D. C. Frick**

* Visual Plumes Consultants, 1541 NW Spring Street, Newport, Oregon, 97365-2344, USA
(E-mail: *wefrick@yahoo.com)*
** Clayhouse Wines, Paso Robles, California, USA
(E-mail: *mekadek@yahoo.com*)

Abstract
Many environmental variables, including beach bacteria concentrations and extreme temperatures, are affected by meteorological conditions. Multivariate Linear Regression (MLR) models prove useful for relating response variables to explanatory variables when modeling average conditions. However, extreme values are often of greater interest; this is true of bacterial concentrations in beach water or freezing temperatures in vineyards. In both instances the effects can be severe to humans or plants. These phenomena are not unrelated as the conditions that lead to extreme temperatures may also be conducive to high bacterial concentration in water. Also, examinations of model results often reveal bias in extreme forecasts. In the vineyard example, the extremes of both low and high temperatures will be underestimated. With beach bacteria, correctly predicted violations of health standards are often overshadowed by a great number of false negative predictions. This work attempts to reduce the tendency to bias extreme predictions to the mean. One technique is to develop response variables that are expressed relative to some reference quantity. For example, it can be shown that in many lowland locations the 850mb temperature is a function of the tropospheric air column thickness. It proves useful to define an 850mb reference temperature based on the regression of these variables. The difference between the surface temperature and the reference temperature becomes the response variable. This approach has the inherent advantage of placing the variable of interest, the overnight low surface temperature, into a seasonal context. Some selected explanatory variables are key parameters in global circulation models, as maintained by the United States National Weather Service. They are readily available and can be shown to be predicted accurately. In addition, to the 1000-500mb thickness and 850mb temperature, other variables include precipitable water, wind, and sky cover. Recommended variables and methods discussed herein help produce accurate forecasts of extreme events.

Keywords
Extreme events; forecast; multiple linear regression; solar radiation; frost; beach bacteria, atmospheric water

INTRODUCTION

At first glance aquatic plume modeling, beach bacteria prediction, and forecasting extreme low temperatures in a vineyard in California might appear to have little in common. However, fresh-water beach discharges or those discharged through outfall structures, such as diffusers, can obviously represent sources of beach fecal bacteria contamination (Comino et al. 2008, 2010). Once in the water, exposure to ambient conditions such as temperature, salinity, and solar radiation can modify bacterial concentrations; intense sunlight in particular is lethal to fecal bacteria. Finally, atmospheric conditions can alter aquatic properties by direct interaction at the turbulent surface interface or through transmission of solar radiation. Frick, Ge, and Zepp (2008) and others have found that atmospheric moisture variables and winds often prove to be valuable explanatory variables in the identification and construction of multiple linear regression (MLR) statistical models (Ge and Frick 2007, Nevers et al. 2007). Moisture variables such as dewpoint temperature often prove to be better explanatory variables than measures of total solar insolation; this appears reasonable because moisture in the atmosphere partially absorbs some of the light frequencies active in bacterial photo decay, significantly modulating the lethal capacity of solar radiation (Frick, Ge, and Zepp 2008).

For health officials and vineyard managers alike, extreme events—elevated bacteria concentrations or killer frosts—are primary concerns. Quite literally such episodes can be killers, for bathers exposed to excessive doses of waterborne pathogens and for grapevines from very cold temperatures that destroy fruit before it can be harvested or late-spring frost damage to new leaves and fruiting bodies. Far from being unrelated, extreme low temperatures and high bacterial concentrations are caused by many of the same basic conditions: low atmospheric moisture content, the absence of clouds, winds, and other factors.

Concerned individuals in both fields must also be aware of a tendency of many numerical and statistical models to underestimate the severity of extreme events. For example, lower low temperatures are forecasted that are systematically too high (and, though not a concern in the vineyard, warmer low temperatures during warm episodes are systematically forecasted too low). Thus there may be substantial periods in the time-series record of bacteria concentration and temperature forecasts that show bias that may persist for several days.

Statistical model building tools, such as Virtual Beach (Zepp et al. 2010; Frick, Ge, and Zepp 2008), are available to identify MLR equations, or models, that fit empirical data well. This is significant as it allows users to develop models that provide "second opinions" that are independent of other sources of information, perhaps official sources like the National Weather Service (NWS) in the United States. For bacteria prediction, these models likely will be the primary means for nowcasting or forecasting concentrations, as other sources for prediction are not available. The object of the custom MLR model is to predict the response variable, the variable of concern.

But, whether the custom MLR models represent a primary means for prediction or not, when true forecasting is involved the explanatory variables may well be forecasted variables themselves. For example, a custom MLR model might use an official estimate of an environmental variable as an explanatory variable. However, using such a variable, itself closely related to the response variable, as an explanatory variable to fit a model will likely transmit some of the bias apparent in the official source that provided the forecast of the explanatory variable in the first place. The second opinion will not be unbiased or independent.

This work examines a few such specific and implicit assumptions and practices that can affect both the performance and independence of custom MLR models. Examples of the genre include cognate variables. Cloudiness is an example of this genre; typically cloud cover is reported in categories while it is forecasted in percent cloud cover. The categories are clear, few, scattered, broken, and overcast, which might be assigned values from 1 to 5 inclusive (as we did for some models). How does this system relate to percentage cloud cover? A regression might be performed to answer that question. But it becomes more complex in the context of predicting overnight low temperatures. What about fog? If it is a radiation fog, perhaps it should be assigned the value zero (a clear sky taken below the condensation temperature near the surface). Meanwhile, an upslope advected fog might be assigned the value 5 or even 6, a very dense overcast. Other peripheral issues involve defining the day's low temperature. Most of the time the low of the day occurs near dawn but sometimes it occurs late the next night. Not differentiating can impair the regression.

Finally, bacteria concentrations and low temperatures appear to have seasonal components. As such, what should be regressed, the absolute value of the low temperature or the difference between the low temperature and some reference temperature? This work supports the latter approach.

THE PASO ROBLES KILLER FROSTS OF 8 AND 9 APRIL 2011

The city of Paso Robles (El Paso del Robles) is situated on the upper Salinas River about halfway between San Francisco and Los Angeles, California, about 180mi (290km) from either city. While often moderated by marine influences, it is separated from the coastline by the Santa Lucia Mountains that rise to about 3000ft (about 1000m) in that region. The Paso Robles airport, the primary point for weather prediction and observation in that region, lies about 17mi (27km) inland (35.63N, 120.69W; elevation 725ft or 221m). The NWS maintains Paso Robles as one of California's climatological stations.

Figure 1. Young grape shoot, in Templeton, California, about 5mi (8km) south of Paso Robles, damaged by the severe late frost of 8-9 April 2011. Note, while many leaves are "burned", the fruiting body (flower cluster) either survived the cold or grew in the week after the event but before the picture was taken.

The region surrounding Paso Robles is a thriving wine growing area with about 180 wineries represented. Red Cedar Vineyard (RC; Clayhouse label) is located from about 10 to 13mi (16 to 21km) east of the airport. It is significant because the authors can access the vineyard's extensive weather records. It lies generally south of the Estrella River, a tributary to the Salinas River. The region surrounding Paso Robles is a thriving wine growing area with about 180 wineries represented.

Figure 2. Severely damaged grape vine (same event as in Figure 1). Templeton, California.

Figures 1 and 2 show grapevines damaged by the killer frost experienced in the Paso Robles central coast grape growing region of California on 8 and 9 April 2011. The severity of this event was not forecasted. Technically a frost was not recorded, the minimum temperatures at the airport being 32F (0C) on both days. However, temperatures below freezing were recorded in Block 27 (of cabernet sauvignon) in the southwest corner of the RC vineyard (Figure 3). Block 27 is significant because it is a slope tract with elevations to about 1400ft (425m) that tends to be warmer at night than other sites in the vineyard. It is also not outfitted for wintertime spray irrigation like some blocks. Spray irrigation is used to help prevent frost damage in some blocks of the vineyard. Of course spray irrigation modifies the local microclimate and can spoil MLR model fitting. In contrast to Block 27, Shandon West (35.64N, 120.49W), at an elevation of about 870ft (265m) in the western part of the vineyard near the banks of the Estrella River, recorded 30.8 and 30.7F (about -0.7C) on the 8th and 9th respectively while, or despite,

Date	ET	Temp	T. Max	T. Min	Temp2	Hum	Dew Pt	Solar	Wind	Gust	W. Dir	Rain
4/11/2011	0.13	50.8	65.5	38.4	56	90	40.2	272	2.8	13.2	68	0
4/10/2011	0.14	48	66.5	33.7	54.2	84	37.1	299	3	12.3	258	0
4/9/2011	0.13	44.7	62.6	30.6	52.5	89	34.8	294	3.3	16.9	255	0
4/8/2011	0.09	39	50.4	29.8	53.5	92	31.3	234	2.8	16	150	0
4/7/2011	0.05	42.3	50.8	32.7	59.3	86	36.8	140	3.6	17.9	257	0.36
4/6/2011	0.17	54.1	73.2	41.2	62.9	78	41.2	302	3.1	16.3	267	0

Figure 3. Temperatures (F) and other parameters recorded in Block 27 from 6-11 April, 2011. The minimum recorded temperatures were 29.8 and 30.6(F) (about -1.2 and -0.8C respectively). Note, Block 27, as a slope tract, tends to have warmer overnight low temperatures than the low-lying blocks, such as Shandon West (not shown). The 0.36in of recorded rain on 7 April was natural, not irrigation water.

recording 2.95in (75mm) of irrigation.

In contrast to actual events, on the morning of 7 April freezing temperatures were not forecasted for the next morning for Paso Robles or for a site 14mi (22km) east of the airport (i.e. near the east boundary of the vineyard). The predicted low temperatures were 36 and 33F respectively. (The NWS website allows users to generate and view spot forecasts for nearby sites simply by clicking on a map.)

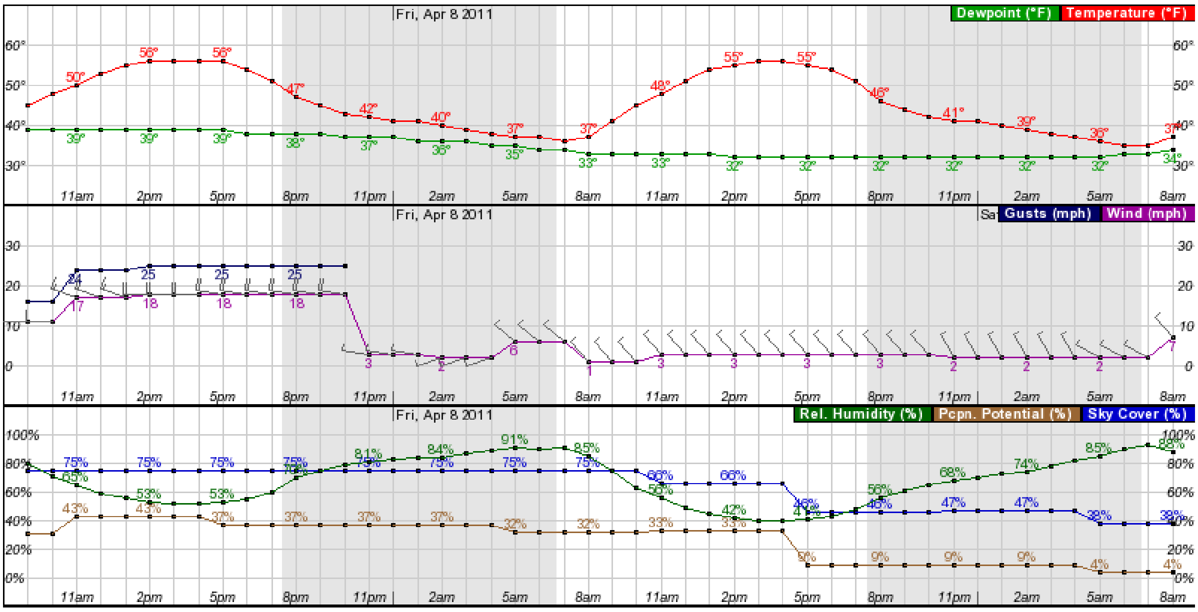

Figure 4. The hourly weather forecast for Paso Robles Airport on the morning of 7 April 2011. Temperatures are reported in degrees Fahrenheit (F), precipitation amounts in inches (in), and wind speeds in miles per hour (mph). Notice that cloudy conditions (75% cloud cover) were predicted to persist throughout the night of 8 April, with light forecasted to persist through about 4pm of the 8th. In fact, clouds dissipated overnight. While clouds reflect solar radiation during the day, they intercept and re-emit terrestrial radiation, serving to limit outgoing energy.

Figure 4 shows the hourly forecasts for 48 hours for Paso Robles beginning at 9am on the morning of 7 April. The overnight low temperatures were forecasted to be 36 and 35F for the mornings of the 8th and 9th respectively. The next day, 8 April, the 24-hr forecast was revised to 34F. On 4 April, four days before they occurred, the senior author emailed the co-author: “I hope to leave for Templeton perhaps Sat or Sun [9 or 10 April] and stay for most of the week. It may be cold the first couple of days. Given the [NWS] charts and prognoses are right (most notably 50% clouds Fri night), [our] Shandon West model predicts 29.4. If it clears up early in the night, it could be even colder.”

On the morning of 4 April the 96-hr low temperature NWS forecast for 8 April for Paso Robles was 39F (3.9C), which was maintained the next day before being lowered to 36F (2.2C) on 6 April. It appeared clear that this system had the potential for producing considerably lower temperatures than were being forecast.

In fact, it did clear up the night of 7-8 April and the conditions were right for a killer frost event.

MLR MODELING, A FIRST LOOK

As the adopted measure of success was to approach NWS in terms of skill (adjusted R^2 values > 70%), forecasting overnight low temperatures well has proved difficult. It was not until late February 2011 and only after some other NWS weather prediction variables were added to the list of potential explanatory variables that this effort began to produce comparable results. A breakthrough appeared to occur after defining a reference temperature based on a regression of the 850mb temperature and the 1000mb to 500mb thickness:

(1) $T_{ref} = -326.5 + 0.06023$ (00Z 24hr forecasted thickness)

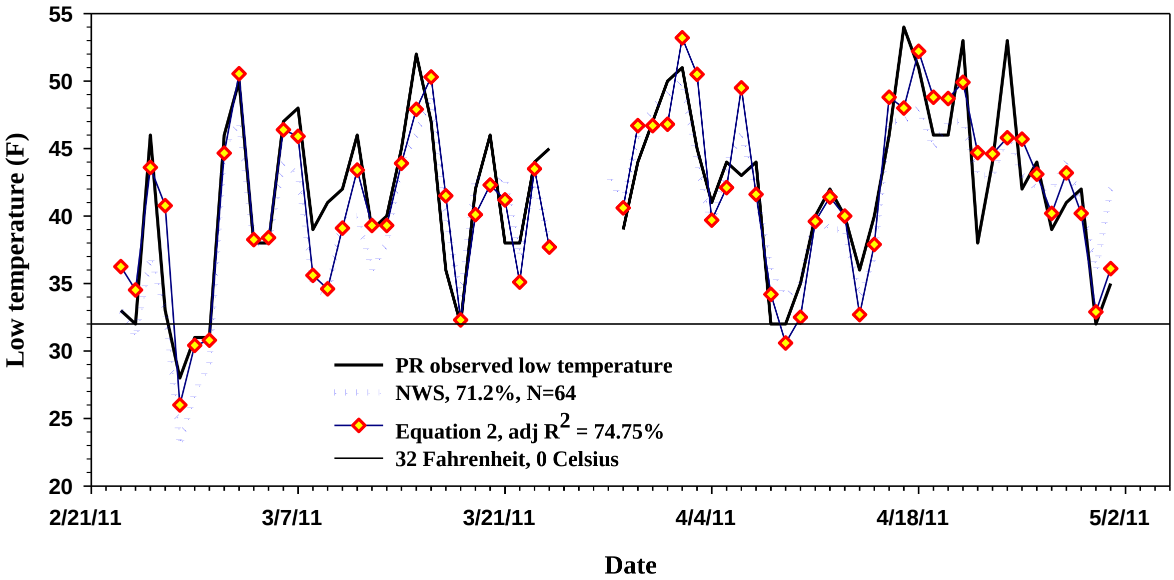

Figure 5. MLR model Eqn. 2 using Tref. N = 64.

The final response variable is $T_{ref} - T$, a difference, not a value on an absolute scale, where T is the observed overnight low temperature in degrees Celsius. As the intent is to forecast low temperatures at least 24 hours in advance, the appropriate independent variable is the 24 hour thickness forecast. Hindcasts of archived observations appeared to confirm that this approach implicitly introduced a seasonal fluctuation in the final forecasted low temperature.

The first of these models (MLR equation) has been maintained since 23 Feb 2011, see Figure 5.

(2) $T = T_{ref} - (-255.92 - 0.9583*cld - 0.3444*NWST - 0.1687*24h2o + 0.501*12th24)$

where T is the forecasted overnight low temperature (in C) and, all ***forecasted 24 hours in advance***, *cld* is the normalized cloud category (approximately 1 to 5, regressed on cloud cover), *NWST* is the NWS low temperature forecast (F), *24h2o* is the atmospheric precipitable water (mm), and *12th24* is the 12Z 1000-500mb thickness (m). The adjusted R2 of this model is 74.7%, fairly similar to the performance of the NWS forecasts (71.2%). (The percentages seesaw as more days are added.)

It is worthwhile to take a few moments to understand the process that produced this MLR equation. The model building tool was the EPA Virtual Beach (VB) model development application (Version 1.0, Frick, Ge, and Zepp 2008). Briefly, VB uses backward elimination to recommend variables for elimination based on minimizing the Mallows Cp value, a parsimony criterion. Each candidate variable increases the Cp value by a unit increment. For a variable to remain among the recommended variables the fit must decrease Cp by more than a unit amount to compensate. Various facilities are available to process the input data, that is, the explanatory variables. For example, VB can check for multi-colinearity, produce interaction terms, and perform transformations. It also can convert speed and direction into two new explanatory variables, the velocity vectors. (However, wind variables have greater uncertainty associated with predicting them, as is evident from their less robust significance values. Scalar variables appear to have an advantage there.)

Equation 2 and other model equations emerged from a process that eliminated many potential explanatory variables. Not surprisingly the NWS low forecast is a robust predictor. However, the other variables are perhaps more interesting in a physical sense. The reference temperature can be understood to modulate the prediction to correct for long-term variations in the response variable (summertime 850mb temperatures tend to be greater than wintertime values). Equation 2 shows that the tropospheric layer thickness may further modulate the overnight temperature. But, emerging from other good predictors such as wind speed, pressure, and even rainfall are precipitable water and clouds, two variables and can directly modulate solar and terrestrial radiation. Thus this purely statistical approach to the problem leads us to a conception that is consistent with our understanding of the physics of the underlying processes.

The results are satisfying, especially considering the critical period of about 7-11 April. The MLR equation predicted the low temperatures during that period better than the NWS, and it did so some days in advance while the official forecast was still forecasting considerably warmer low temperatures.

The problem with this model is that over half of the variance is due to the strongest explanatory variable, the NWS overnight low temperature forecast. Thus the other terms represent a perturbation of the NWS forecast. In this instance it turned out well but in general the Eqn. 2 will transmit some of the bias, if present, of the NWS forecast. For example, in the period 7-13 March both models predicted consistently too low, albeit the MLR model corrected somewhat. The model is not a very independent opinion.

INDEPENDENT SECOND OPINIONS

By early March the problem of model dependence became clearer and the search began for robust models that would ***not*** include NWS low temperature predictions or closely related variables (like the corresponding overnight dewpoint temperature prediction) as explanatory variables. Some variables remain important but others that join the list are the 850mb temperature and the 500mb height forecasts. Depending on the length of the fit and periods of time used, one of the wind components sometimes passes the Cp criterion, usually involving an asymmetric transformation that can be easily performed in VB but is harder to implement in the master spreadsheet.

Fitting to vineyard low temperature data (the Shandon West block) instead of to Paso Robles temperatures produced a model fit that was independent Paso Robles observations, thus allowing a longer test of 24-hour forecasts to be made for Paso Robles. When the plants are dormant there is no irrigation to prevent frost damage or skew the fit. Archived vineyard data were analyzed that showed that the airport lows were approximately 5F (about 2.8C) warmer on average than Shandon West. The model fit equation produced on 23 February was

$$(3)\ T_{PR} = (T_{ref} - 9.793 + 0.08224\,cld - 1.0086\,T_{ref} + 0.4681 * h2o + 0.2961 * T_{0Z850})1.8 + 32 + 5$$

where T_{PR} is the forecasted Paso Robles overnight low temperature. All the remaining variables represent 24 hour forecasted values. They are the reference temperature, T_{ref} , the percent cloud cover, *cld*, the precipitable water, *h2o*, and the 00Z 850mb temperature, T_{0Z850}.

The adjusted R^2 value for a period of 100 days from 21 Jan to 1 May 2011, with four days with missing data excluded (N = 97), is 71.9%; for comparison, the corresponding NWS value is 74.8%. However, the technique of adjusting the Eqn. 3 forecast by averaging the NWS and Eqn. 3 forecasts

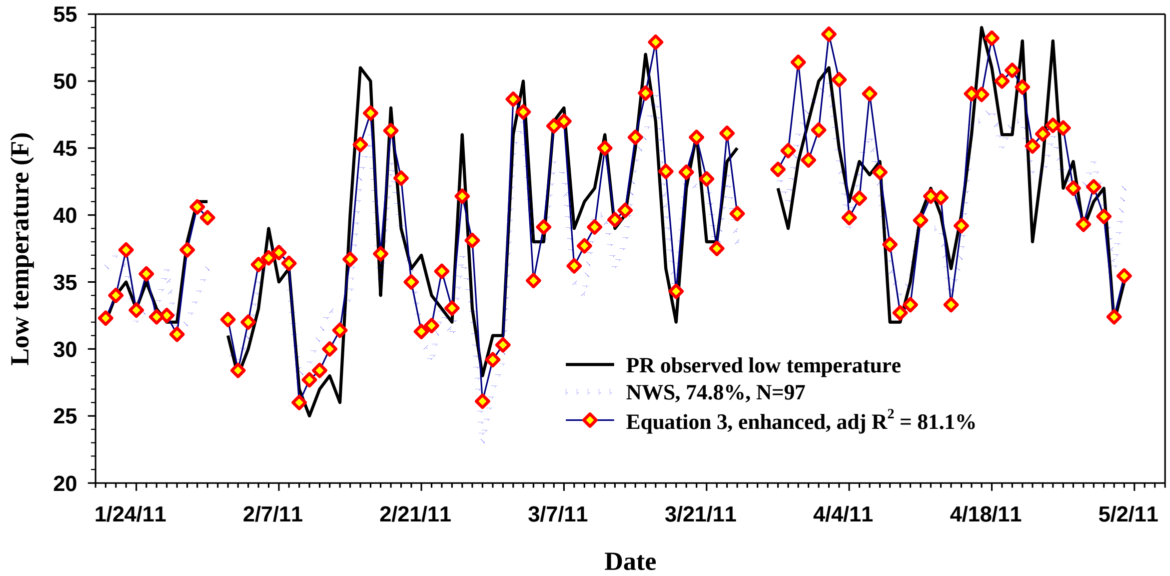

Figure 6. 24-hour low temperature forecasts for Paso Robles based on using Shandon West overnight low temperature in the response variable differential for fitting the model equation (Eqn. 3, with averaging). **Concerning the frost event, *the 8 Apr 24-hr forecast was for 75% morning cloud cover, greatly overestimating clouds that cleared off. At 25% cloud cover Eqn. 3 would have yielded a 30.4F forecast.***

when the latter differs by more than 4F from the former yields adjusted R^2 value of 81.1%. The latter comparison is shown in Figure 6.

Another perspective is offered by comparing forecasts in observed vs. predicted format (Figure 7). Panel (a) shows the raw forecasts using the Paso Robles model fit to Shandon West observations, making it independent of Paso Robles observations; (b) the same forecasts with some “enhanced” using the difference dependent averaging technique for raw forecasts differing by more than 4F from NWS forecasts; (c) the corresponding NWS 24-hour overnight low temperature forecasts. In (c), the hypothesis that cold low temperatures tend to be forecasted high while warm lows tend to be forecasted low appears to be supported in this instance. (N = 96, not changed to include 1 May.)

Recalling that a primary project goal was to provide better extreme low temperature forecasts, the averaging adjustment to Eqn. 3 does not affect any instances of observed temperatures below freezing. In this range of values there is a lesser set of dominant weather phenomena than in instances of warmer overnight low temperatures. The coldest temperatures tend to occur when the air is clear and dry and winds are light, there is less model degradation from hard to forecast cloud and wind effects. The explanatory variables in the parsimonious model well represent these

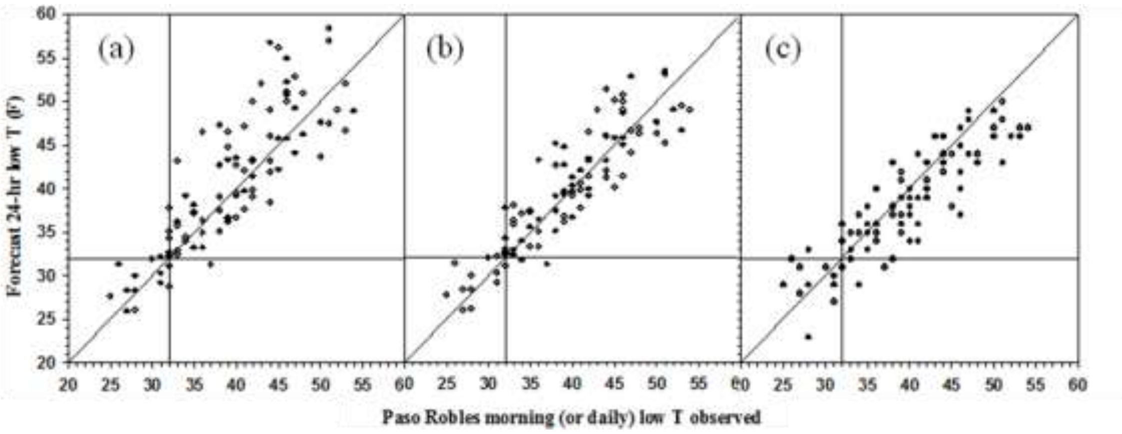

Figure 7. Predicted versus observed plots. (a) Raw Shandon-West PR forecasts; (b) Shandon-West PR forecasts after applying the average forecast technique for differences between NWS and the MLR model greater than 4F; and (c) NWS forecasts; note the bias at high and low values.

conditions. In contrast, the warmer lows occur under a multitude of conditions, including marine and continental conditions, high and low winds from many sectors, a range of cloudiness and radiation and advection fogs not forecasted as well by a parsimonious model of only four explanatory variables, not including winds and other phenomena.

DISCUSSION AND CONCLUSIONS

It is possible to develop MLR models to forecast environmental variables accurately, independent of other parallel forecasts of the same variables. This is at least true for overnight low temperatures and encourages efforts to prove that similar results may be obtained for other variables as well. It might be kept in mind that the data collection effort supporting this work involved transcribing digital data and interpreting and interpolating atmospheric charts, tasks that are subject to some error that surely degrades the performance statistics somewhat. Hence, obtaining independent models, essentially second opinions, which achieve similar performance statistics as NWS forecasts is very satisfying.

Methods and techniques developed and used herein may prove useful in other efforts. In particular, beach bacteria concentrations are modulated by atmospheric conditions much as are overnight low temperatures. Low atmospheric moisture (precipitable water), clear skies (less than 10 percent cloud cover), and winds can modulate beach bacteria concentrations much as they modulate overnight low temperatures. Regressing on a differential, i.e. on a departure from a reference value, proved to be the single most important change in methodology that made this customized effort competitive with NWS forecasts, and ultimately made an independent second opinion possible.

It is not immediately apparent how the reference value technique might influence efforts to predict beach bacteria concentrations. However, as forecasting services begin to provide beach bacteria forecasts, defining reference values and providing independent second opinions may emerge in that field as well. At a minimum, it may prove to be that custom MLR models might provide forecasts of important beach bacteria nowcasting variables, notably water turbidity (Francy and Darner 2006). An improved data collection methodology, based on automatically acquiring digital data and perhaps persuading NWS to provide more chart contours of low moisture variables, provides optimism that more accurate extreme value forecasts may be possible in the future.

ACKNOWLEDGMENTS

There are more individuals and organizations who contributed to this work than can be fully acknowledged. We thank Red Cedar Vineyard and Clayhouse Wines for providing vineyard meteorological data; David Schwab and the NOAA Ocean and Human Health Initiative people for their constant support; Zhongfu Ge for his role in developing Virtual Beach (VB; Version 1.0) and always erudite advice; Donna Francy for providing conceptual seed corn; Ana Paula Comino and Marcelo Scagliola for their early adoption and subsequent use of VB 1.0 and joyful collaboration; Richard Whitman for our Great Lakes "ocean" cruises. Marirosa Molina and Candida West for stalwart and inspiring support of Virtual Beach; John Wathen for his invitation to the EPA Beaches Conference in Miami, March 2011, stimulating the rapid improvement of techniques described in this paper; and David Rockwell for "spiritual" insight into bacteria, weather forecasting, and life.

EXTENDED REFERENCES

Baumgartner, D.J., W.E. Frick, and P.J.W. Roberts, 1994. Dilution models for effluent discharges (Third Edition). Pacific Ecosystems Branch, ERL-N, Newport, OR. EPA/600/R-94/086, Jun 1994, 189pp.

Comino, A.P., M. Scagliola, Z. Ge, and W. Frick 2010. Integrated coastal management strategy of Mar del Plata city and the use of Virtual Beach as a management tool. Proceedings of MWWD 2010, the 6th International Conference on Marine Waste Water Disposal and Marine Environment (www.mwwd.org), Langkawi, Malaysia, 25-29 Oct 2010.

Comino, A.P., M. Scagliola, Z. Ge, W. Frick and L. Patat 2008. Beach Water Quality of Mar del Plata (Argentina) previous to outfall construction: ten years of bacterial control and Virtual Beach model prediction testing. Proceedings of MWWD 2008, the 5th International Conference on Marine Waste Water Disposal and Marine Environment (www.mwwd.org), Cavtat/ Dubrovnik, Croatia, 27-31 Oct 2008

Francy, D. S., R.A., Darner. Procedures for Developing Models to Predict Exceedances of Recreational Water-Quality Standards at Coastal Beaches, U.S. Geological Survey: Columbus, OH, 2006.

Frick, W.E. and D.C. Frick 2011. Methodologies to complement the direct application of MLR model development. Oral presentation. EPA National Beaches Conference, 14-17 Mar 2011, Miami, Florida.

Frick W.E., Z. Ge, and R.G. Zepp 2008. Nowcasting and forecasting concentrations of biological contaminants at beaches: A feasibility and case study. Environmental Science & Technology. American Chemical Society, Washington, DC, 42(13):4818-4824, (2008)

Frick, W.E. and Z. Ge 2006. Bacteria, beaches and swimmable waters: Introducing Virtual Beach. MWWD 2006 4th International Conference on Marine Waste Water Disposal and Marine Environment, Antalya, Turkey, Nov. 6 10, 2006. 23 pp.

Frick, W.E., P.J.W. Roberts, L.R. Davis, J. Keyes, D.J. Baumgartner, and K.P. George, 2003. Dilution models for effluent discharges, 4th Edition (Visual Plumes). EPA/600/R-03/025. USEPA, Athens, Georgia

Frick, W.E., C.A. Sproul, and D. Stuart, 1997. Bacterial impacts of ocean outfalls: legal challenges. Journal of Environmental Engineering, 123, 2, pp. 191-196

Ge, Z. and W.E. Frick 2007. Some statistical issues related to multiple linear regression modeling of beach bacteria concentrations. Environmental Research. 103, 358-364

Nevers, M.B., R.L. Whitman, W.E. Frick, and Z. Ge 2007. Interaction and influence of two creeks on Escherichia coli concentrations of nearby beaches: exploration of predictability and mechanisms. Journal of Environmental Quality, 36, 1338-1345

Sigleo, A.C., W.E. Frick, and L. Prieto 2010. Alder cover affects stream water nitrate: comparison of two Oregon watersheds. Northwest Science 84:336-350.

Zepp, R. G., M. Cyterski, R. Parmar, K. Wolfe, E.M. White, and M. Molina 2010. Predictive Modeling at Beaches, Volume II: Predictive Tools for Beach NotificationU.S. Environmental Protection Agency, National Exposure Research Laboratory, EPA-600-R-10-176, Nov 22, 2010