SCIENCE APPLICATIONS INTERNATIONAL CORPORATION

The Economic Benefit of Incorporating Weather and Climate Forecasts Into Western Energy Production Management

Deliverable 4—California ISO: Ensemble Forecast Modeling

Submitted To:

Scripps Institution of Oceanography (SIO)
La Jolla, CA

May 2004

SAIC
An Employee-Owned Company

FINAL REPORT

The Economic Benefit of Incorporating Weather and Climate Forecasts into Western Energy Production Management

Deliverable 4 -- California ISO: Ensemble Forecast Modeling

May 2004

Completed By Science Applications International Corporation (SAIC)

This Page Intentionally Left Blank

Acknowledgements

SAIC Team

Dr. Mary G. Altalo

Todd D. Davis

Monica S. Hale

Dr. Leonard A. Smith

Scripps Institution of Oceanography Contributors

Dr. David Pierce

Dr. Alexander Gershunov

Mary Tyree

Dr. Anne Steinemann

This Page Intentionally Left Blank

SAIC An Employee-Owned Company

Table of Contents

Page

List of Figures and Tables

Page

Figures

Tables

Abbreviations

BS	Brier Score
CDF	cumulative distribution function
CPUC	California Public Utilities Commission
ENSO	El Niño/Southern Oscillation
MSLP	mean sea level pressure
NWP	numerical weather prediction
PF1	Probability Forecast 1
RMS	root-mean-square
RMSE	Root Mean-Square Error
SDG&E	San Diego Gas and Electric Company
SONGS	San Onofre Nuclear Generating Station
SLP	sea level pressures
SPP	Statewide Pricing Pilot
TOU	time of use
WMO	World Meteorological Organization

This Page Intentionally Left Blank

Chapter 1: Improved Prediction of "Tariff Days" To Support SDG&E's Price Responsive Load Response Program and The Economic Benefit

1. Background

San Diego Gas and Electric Company (hereafter SDG&E) is a part of Sempra Energy Utilities, the umbrella for Sempra Energy's regulated energy distribution business units. Sempra Energy (NYSE: SRE) is a Fortune 500 energy services holding company based in San Diego. SDG&E is a combination natural gas and electric utility that serves much of San Diego County and a small portion of Orange County, California. SDG&E has $1.5 billion electric revenues, of which, $649 million comes from the residential market, $632 million from the commercial market and $161 million from the large commercial/industrial market. There is little industrial load in San Diego County; about 1.1 million residential electric customers; 133,000 commercial; and 427 large commercial/industrial customers. Of the total SDG&E generating assets, 3,609,624 comes from nuclear power at San Onofre Nuclear Generating Station (SONGS). SDG&E has about 1,504 pole miles of transmission. **Figure 1-1** presents an illustration of the monthly peak load for SDG&E.

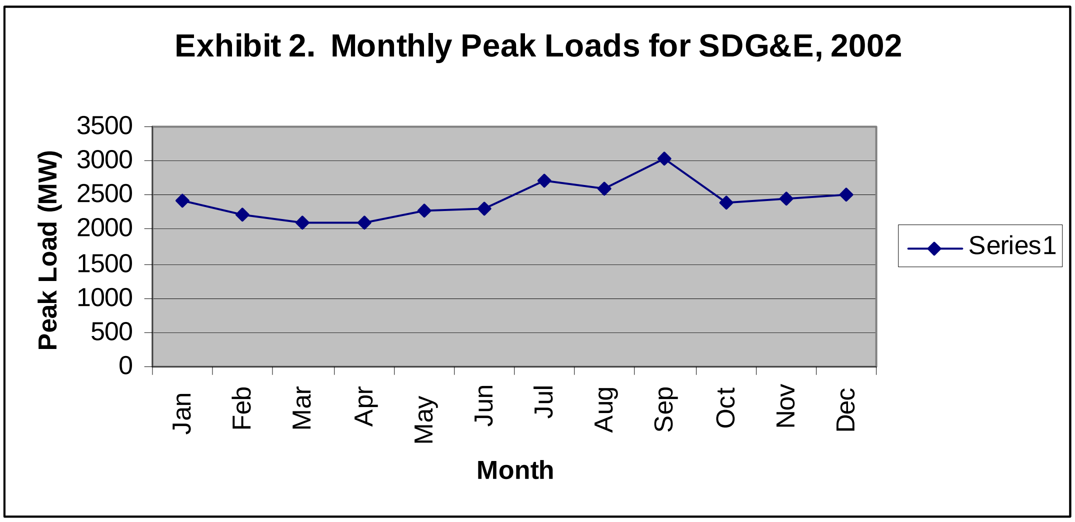


Figure 1-1. SDG&E Monthly Peak Loads, 2002

Service Territory

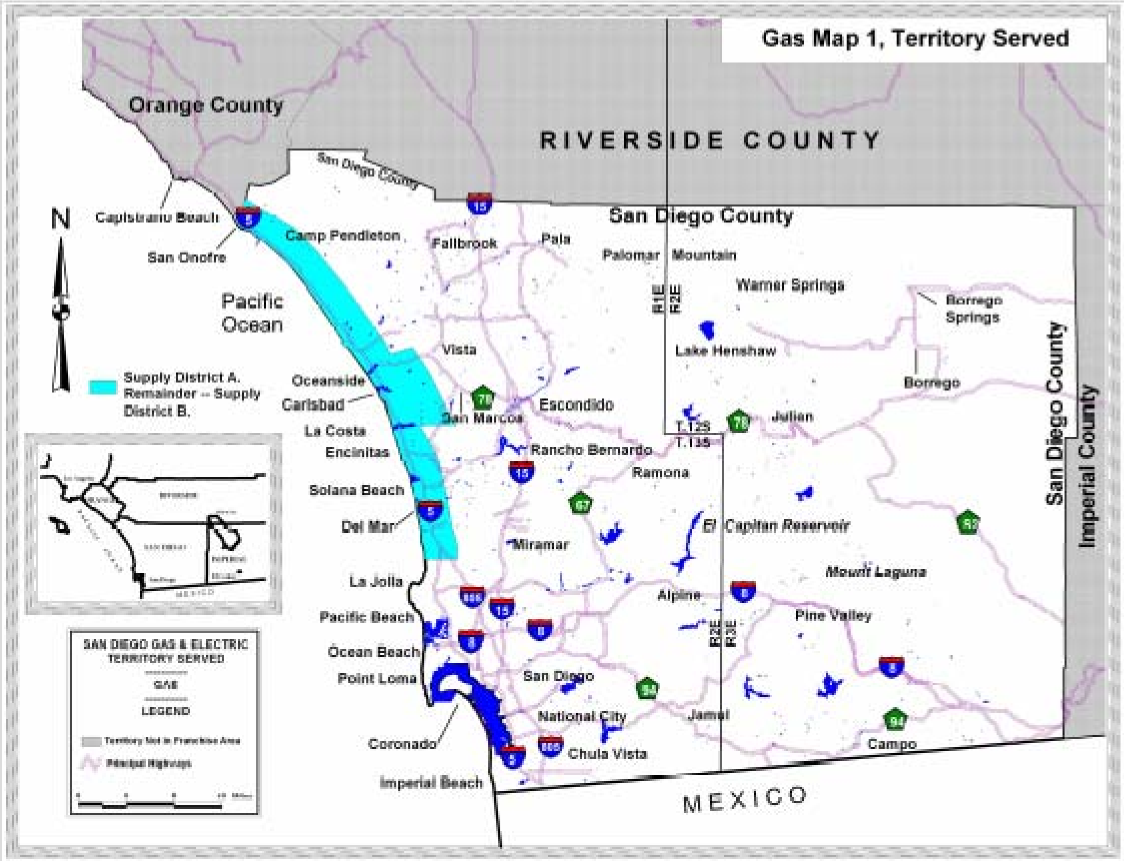


Figure 1-2. Service Territory of SDG&E

The Peak Load Reduction Program

SDG&E, along with other statewide utilities, are completing a pilot demonstration project to evaluate the effectiveness of tariff events on the residential load. Over 2,500 customers statewide are participating in this project with four different rate structures applicable: variable, fixed hours, time of use (TOU) rate, and information only (IO) programs. In San Diego, only the variable rate treatment applies. Whereas in the Bay area, fixed, TOU and IO programs apply. These statewide pricing and load control incentives need accurate forecasts of warmer than normal temperatures on a 3-7 day-ahead basis. This information is used to schedule and dispatch load tariff events and to evaluate the effectiveness of these events in developing elasticity results for the system. Presently they are using temperature alone. In many cases, it is the humidity that contributes to the greatest error. Residential customers are issued a signal that day or a few days ahead and it is up to them as to how to reduce load during these period when higher electricity rates (5 to 10 times higher) go into effect. The threshold temperature of 91 degrees Fahrenheit at Miramar triggers the "commercial" customers automatic rate increase. Recently, the Public Utilities Commission (PUC) ordered tariffs for commercial customers with reduction targets of 30 MW to be signed up for this year (2004) and a reduction of 80MW projected for next year (2005). The residential tariffs have no set targets as at the time of writing (April 2004). Based on the outcome of these pilot programs, the "real rate" of response residential electric users will be established in the coming year. SDG&E needs to develop skill in "calling" these weather-related events, and influencing customer behavior. While the current pilot is for summer peak temperatures, a similar process could be developed for consecutive cold days in winter.

The Statewide Pricing Pilot (SPP) study was a result of a California Public Utilities Commission (CPUC) Rulemaking issued June 2002 (R.02-06-001). The purpose of the rulemaking is to formulate comprehensive policies to develop demand flexibility. One of the primary objectives is to provide options for customers to respond to dynamic retail prices. The pilot has various tariffs that typically have a critical peak period price. The critical peak price is overlaid on a time of use pricing structure and the critical peak price is in effect during critical peak events. The tariffs allow for 15 critical peak days (12 summer and 3 winter). The SPP Dispatch Team calls or "triggers" the events. Most events are triggered on "quasi" peak days, which are generally warmer than average days. The team must rely on accurate 3-4 day weather forecasts to facilitate declaring an event. Issues surrounding inaccurate weather forecasts are: 1. Loads are not available for demand reduction (too cool, nothing to reduce) – they waste one of only 12 events for load analysis. 2. Loads are high, system is constrained and/or peaking and they do not call an event based on an inaccurate weather forecast (too low) – as they miss out on a good estimate of load reduction and future potential for load reductions.

Costs to the company are difficult to estimate: currently (April 2004) SDG&E are concerned with formulating a good econometric model to estimate price elasticities. This may impact full-scale rollout of new tariffs. In future there may be demand reductions that would be known and only obtained when hot weather and peak loads occur. Normal weather will not yield the same results; the difference between the two could result in a high financial transaction cost to the company.

The Case Study Demonstration Project

As noted above, SDG&E participates in a statewide load control experiment. This increases price signals to customers thereby giving customers an incentive to shed load during critical peak times of up to 15 periods a year (12 summer and 3 winter). The experiment has been "hit or miss" when applying the current weather forecast information to decisions calling for a critical peak-pricing, load-shedding event. The scheduling of the critical peak price activity often occurs when loads are not maximized or the forecast does not indicate a peak load event and one, in fact, occurs. A large part of the load variation in this region is driven by weather variability, so it is desirable to call critical peak price events during times when there are maximum cooling or heating requirements. This has been difficult to achieve—less than a 20% match occurred during the summer of 2003. As mentioned above, the state has allowed for only 15 events per year and these can only be used during a fixed time period (events can be called during a 5-hour window—from 2 p.m. to 7 p.m. on non-holiday weekdays).

This is similar to mismatching hydro plant use when periods of precipitation are expected – but does not occur. Meanwhile the commitment to run hydro generators may have been made when it may not be optimal to do so.

This project's aim is to help SDG&E better forecast with a 3-day lead-time, periods when higher or lower temperatures occur that signal a need for a load management event. Better coordination between expected weather conditions and the best periods for load control can then occur. Scripps Institute of Oceanography (SIO) provided the short term and intermediate forecast for SDG&E program managers and SDG&E will coordinate with the state concerning the timing for invoking a load control event. Also, better predicting these events would be helpful in planning more operational support when the power system is being stretched for meeting peak loads.

A decision system will be provided to SDG&E that supports the timely decision time frame for when a load control event should occur, due to weather conditions. SDG&E will run and evaluate the decision they should make regarding when to initiate a load control event. The decision system may be an Excel spreadsheet where SDG&E inputs such variables as forecast temperatures and humidities at Miramar, ENSO indices, and when previous tariffs were called. It will then give SDG&E information on the likelihood of potential upcoming tariff events that they can use. An important part of the SIO work will be to characterize the skill of this forecast system, so that SDG&E will be able to do a correct pricing and time estimate of the product. Essentially, SIO provided SDG&E a decision system that will evaluate historical and current weather data and forecasts, as well as the number of events already called in a particular season, and give a four-day outlook for upcoming tariff event scheduling.

2. Case Study Objectives

The purpose of this project was to use weather and climate information to help SDG&E better forecast load management events with a 3-day lead-time to allow for improved decision-making utilizing weather conditions to inform the optimum periods for load control. Better prediction of weather events is helpful when planning the operational support needed when the power system is being stretched to meet peak loads.

This project involved SIO providing SDG&E program managers with 3-day lead forecasts of weather variables directed to the critical peak pricing tariffs. SDG&E identified the factors determining the timing for invoking a load control event. The objectives of the case study included:

1. To make optimal use of available weather and climate data in forecasting the 12 summer days for calling load management control events to support the SDG&E pricing experiment program.
2. To rigorously characterize the skill of the weather component of the tariff forecast system, so that SDG&E has the information required to properly evaluate the pricing and cost-effectiveness of the program.
3. To estimate the economic value of the improved use of weather forecast information as it applies to making informed decisions on when best to call a critical peak-pricing event.

3. Approach

The approach used to complete this case study included the following tasks:

1. Integration of the SIO weather forecasting subtasks with SDG&E and SAIC subtasks.
2. A detailed review, or base case assessment, of the past year's (2003) load control events and the actual weather conditions extant at the time.
3. The analysis included identifying the following:
   a. The research questions to be answered, time scale, and desired accuracy of the forecasts.
   b. Identification of the models, data, and analysis methods used to address the major research questions.
   c. Weather forecast data used.
   d. Description of how SDG&E can incorporate the SIO weather forecast and how this is applied to deciding on invoking price-induced load control events.
   e. SAIC and SDG&E identified the economic value of from more accurately linking load control events to ideal weather conditions.

4. Analysis and Results

Background: Load and Error Analysis of Event Days

Figure 1-3 shows the average SDG&E summertime weekday load over the hours 10 am to 4 pm, i.e., the average summer daytime load is X MW (note: in the rescinded version of this report, the values have been normalized to the average summer afternoon).

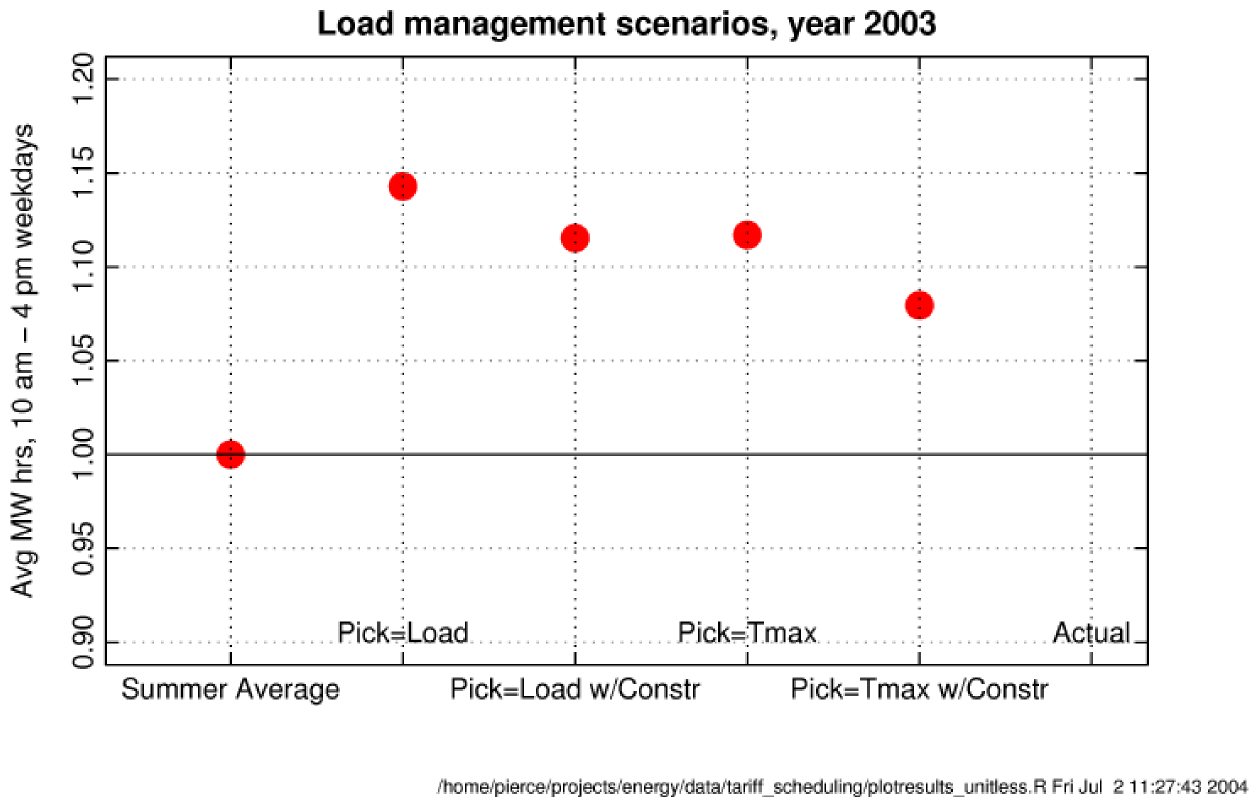


Figure 1-3. Different Load Management Scenarios, 2003

The 12 highest load days during 2003, showed an average daily load 14% higher than average. Note that this does not include any PUC constraints (number of events/week, number of events/month). Also, this is carried out on the basis of load values, not temperature values. Selecting the top 12 hottest days gives a different average load due to the imperfect correlation between temperature and load.

If the PUC-mandated constraints are incorporated (5/month, 2/week maximum), then the average load for those days, selecting on the basis of the load values, is 12% higher than an average summer day.

Selecting temperature values (rather than load, as before) and with no constraints, the average load of the top 12 days is 12% higher than average.

Selecting temperatures with constraints, the average of the top 12 days is 8% higher than average.

The actual average load of the days SDG&E selected in 2003 was XX.

Figure 1-4 shows the growth in electricity use over the period 1977 to 2003 (normalized in the rescinded version). To study the year-to-year variability of the peak load days, it is important to look at a longer timespan than just 2003, as figure 1.3 shows. To do this, we normalize by the trend line over the period 1990-2003 (shown in red in **figure 1-4**). This will allow us to calculate the effect of selecting tariff days over this entire period.

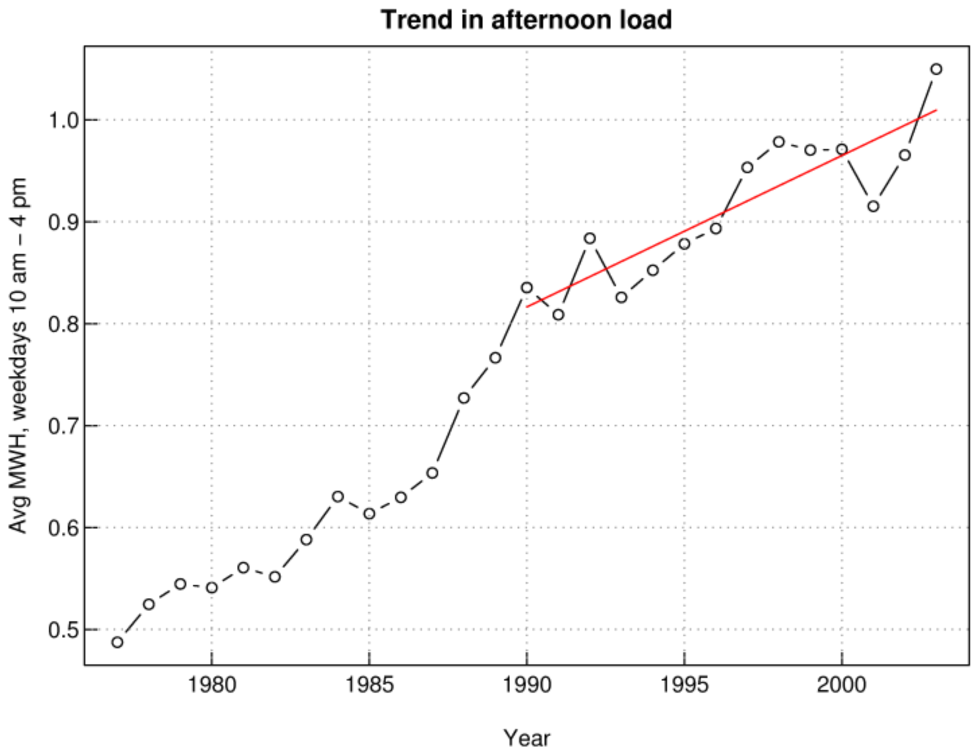


Figure 1-4. Historic Trends in Afternoon Load

Figure 1-5 shows the year-to-year demand level normalized to the red trend line shown in the previous graph. So, for example, 1992 used 105% of the average electrical load on summer days than expected from the trend, while 2001 used 94%.

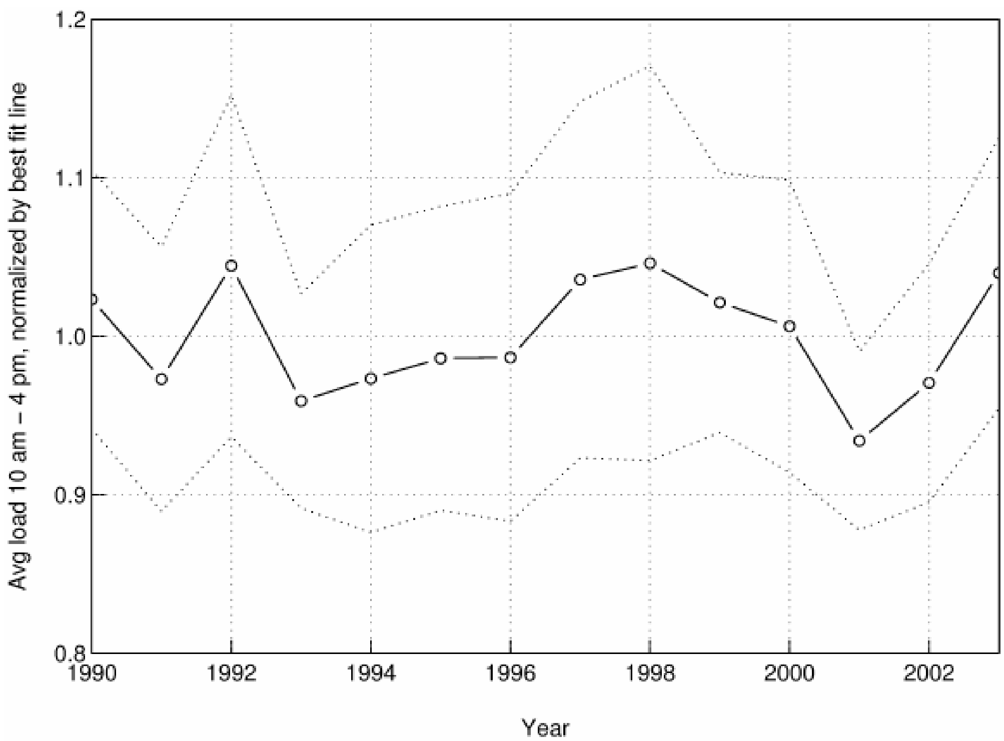


Figure 1-5. Normalized Demand Level

In reviewing the same kind of numbers as before for the average summer day load -- but normalized, one can look at all years from 1990 to 2003. The average summer day load will be exactly 1.0 in these plots.

Figure 1-6 shows how much more the load is than the average summer day when you pick the top 12 events based on either the observed load (points marked "load") or based on the observed temperature (points marked "temperature"). Also shown is the actual value SDG&E picked for 2003.

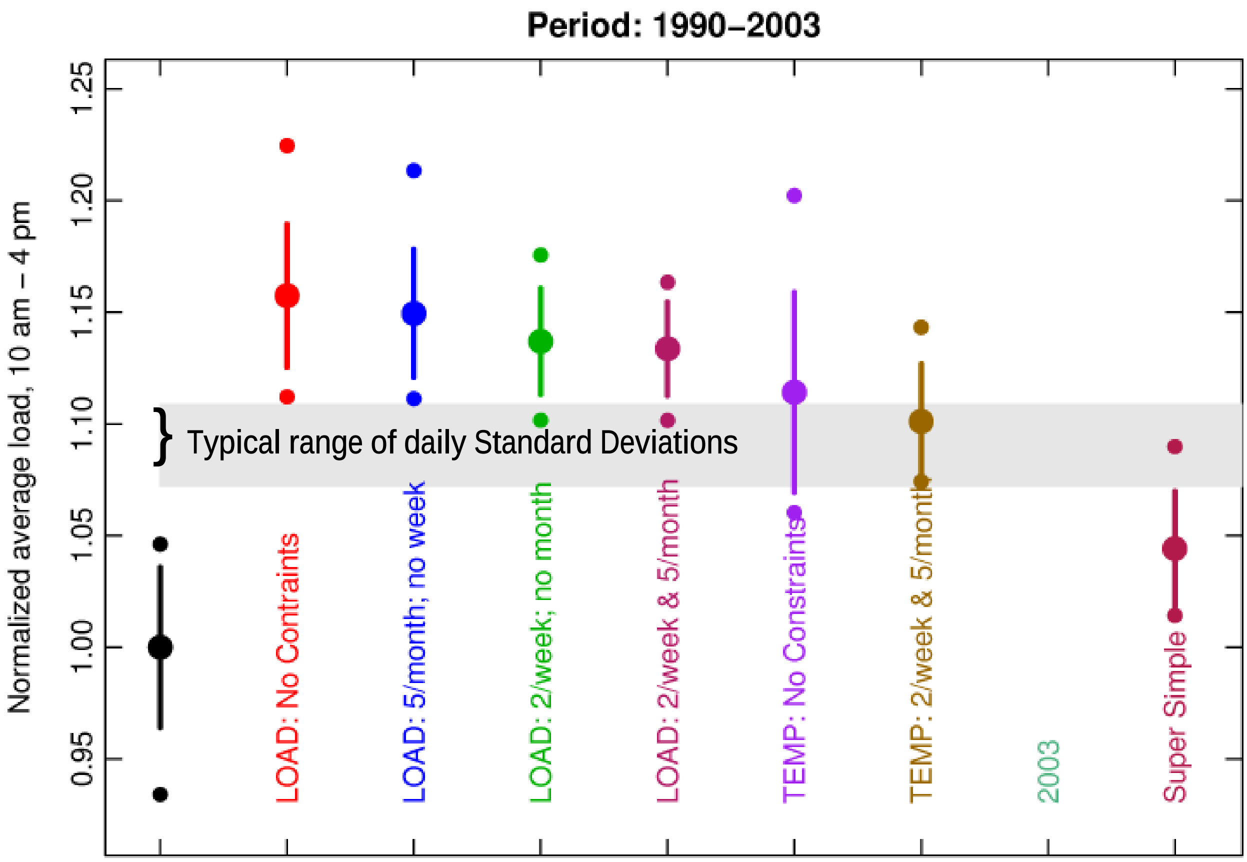


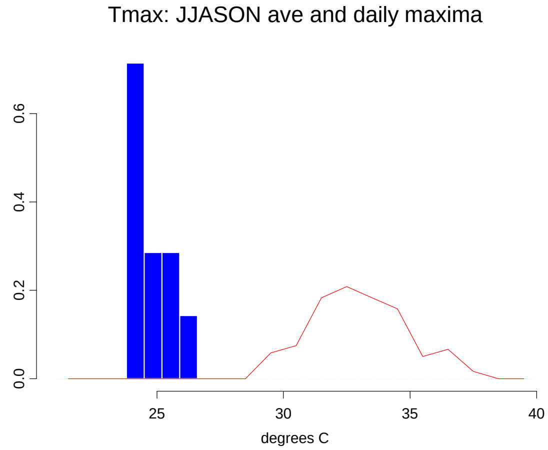
Figure 1-6. Normalized Average Load on tariff days for various methods selecting tariff days: based on observed load ("LOAD") or temperature ("TEMP"), and with or without PUC-mandated constraints. Also shown are the results from the "super simple" statistical forecasting scheme.

The relatively low average load of the days actually picked in 2003 suggests that a straightforward scheme could do better. One such scheme developed and evaluated here is the so-called "super simple scheme," and works as follows. If today is > 26° C (79° F) and is 0.5° C (0.9° F) warmer than yesterday, then call a tariff day 3 days from now (i.e., if criterion satisfied on Friday, call Monday as a tariff day). **Figure 1-6** shows that this scheme picks days with an average load of 6% greater than a typical summer day. Since the maximum that can be obtained based on perfect knowledge of temperatures is only 10%, and the actual performance in 2003 was X%, it can be seen that the super simple scheme both does much better than the method used to pick days in 2003, and recovers a decent part of the maximum attainable by any temperature prediction scheme. Therefore it is both useful and valuable in setting a practical minimum to the benefits of a program that selects days based upon temperature forecasts.

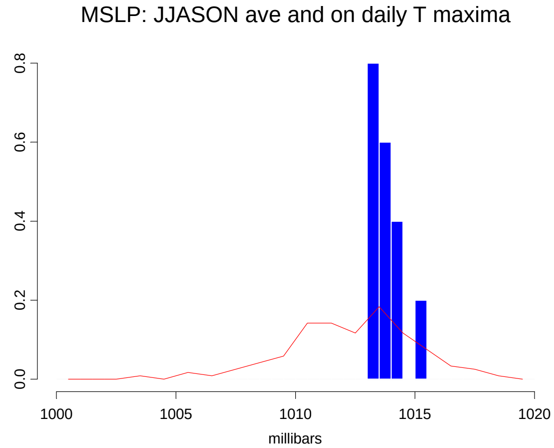
Figure 1-7 shows Tariff Days' temperatures and sea level pressures (SLP) in the context of their seasonal values. Tariff Days' temperatures are warm outliers as per definition, while Tariff Days' SLP, although tending to be somewhat lower than seasonal values on average, can also be close to or even higher than seasonal averages. This means that local SLP probably cannot be used to predict Tariff Days (see **figure 1.10**).

Weather Variations Tied To Event Days: Historical Analysis

Tariff Day season defined as June–November



PDF of seasonal average Tmax (blue) and Tariff average Tmax (red) over the 10-year record



PDF of seasonal average Seal Level Pressure and MSLP averaged over Tariff Days

Figure 1-7. Tariff Days At Miramar Marine Corps Air Station (KNKX)

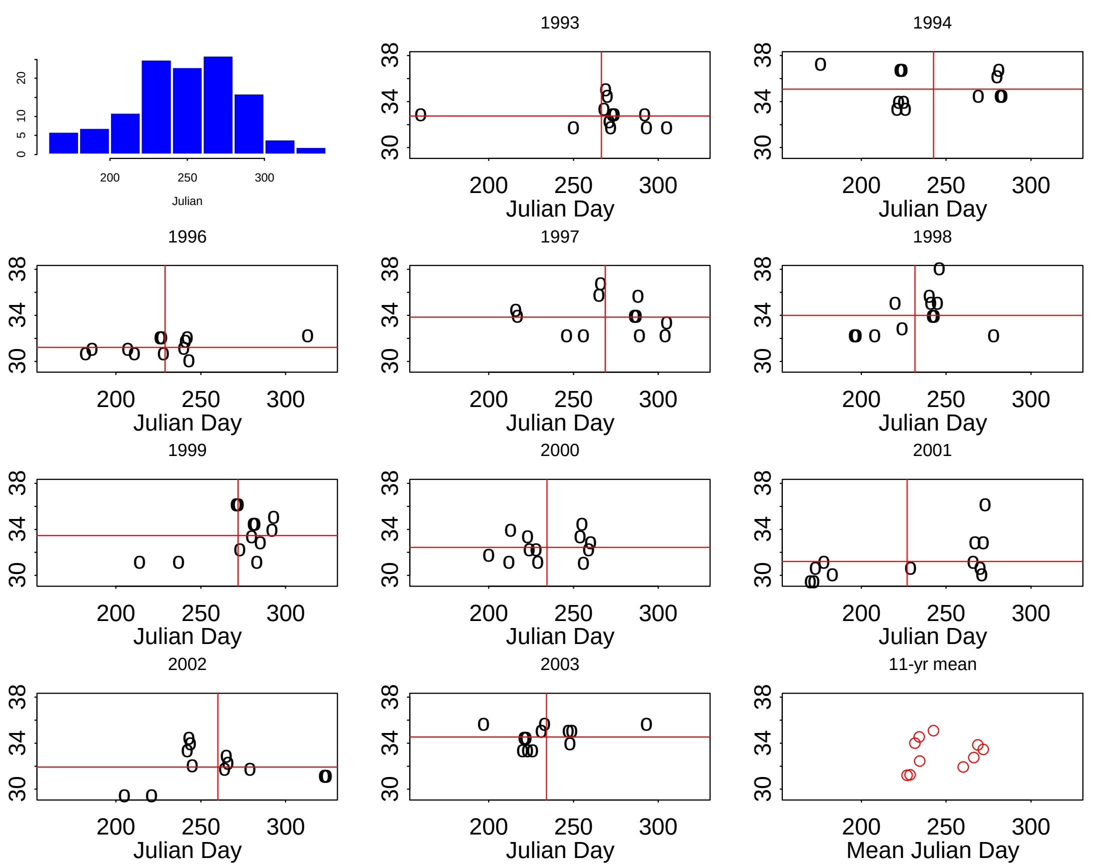


Figure 1-8. Tariff Days' Timing and Intensity by Season along with Average Values

Figure 1-8 shows Tariff Days' timing and intensity by season along with average values. It is clear that the average timing and intensity of Tariff Days in a season are unrelated. It is also clear that Tariff Days tend to occur in clumps and that what may be defined as a Tariff Day in one season may not be in another (compare for example 1994 and 2002). This suggests that the definition of “Tariff Day” is not optimal for seasonal analyses, as it varies with each Tariff Day season and it is impossible to know what the temperature threshold for Tariff Day is until the season is over.

Figure 1-9 shows that Tariff Days tend to occur in runs of consecutive hot days (i.e. clumps). This persistence, although artificially lowered by the arbitrary definition of the temperature threshold defining Tariff Days for each season, is nevertheless strong, suggesting that it is possible with simple persistence to predict Tariff Days when a Tariff Day run (clump) has started. In other words, the real challenge is to predict the first Tariff Day in a run of Tariff Days.

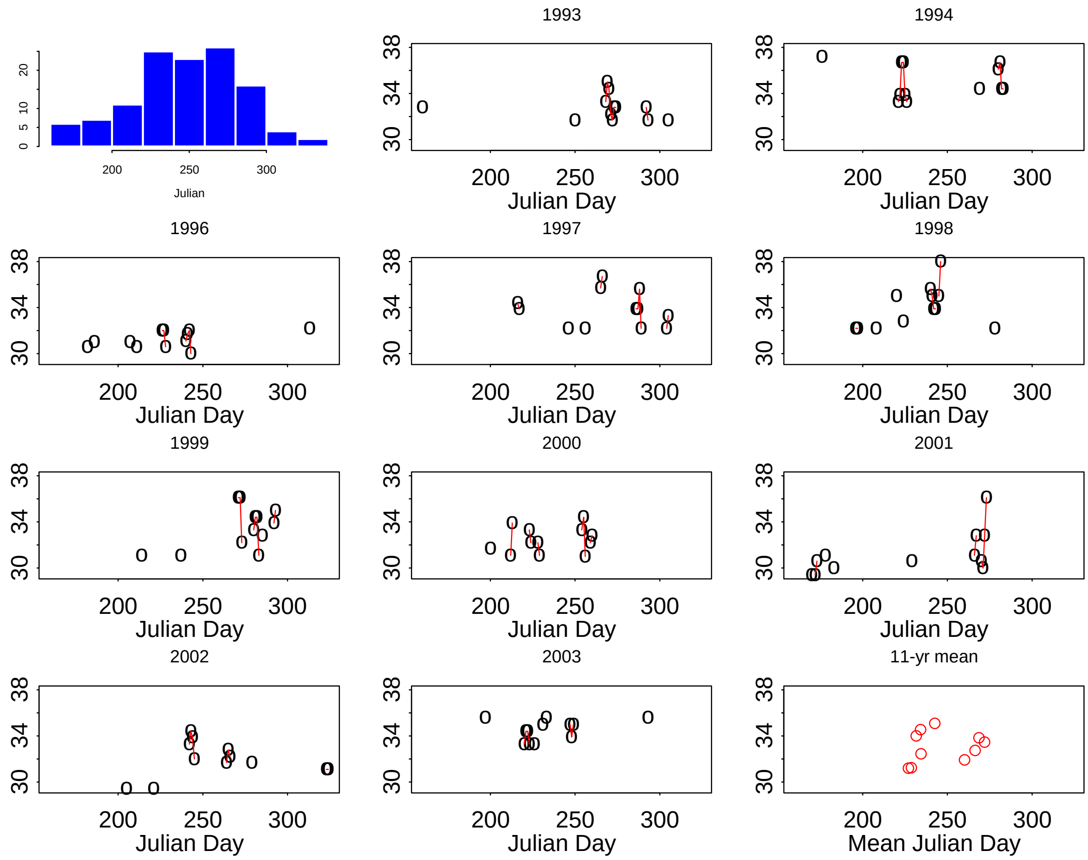


Figure 1-9. Tariff Day Timing-Intensity Plots.
Consecutive Tariff Days are connected with red segments.

Figure 1-10 shows the SLP field (colors) as well as local temperature and SLP (bottom curves) evolution leading up to the first day of a Tariff Day run. Figure 1-7 shows that local SLP has no particular signal on a Tariff Day. Comparing the field and local evolutions of SLP in these two figures and the two figures below (Figure 1-11), neither the spatial pattern of SLP, nor the local values evolve in a consistent way leading up to the first day of a Tariff Day run suggesting large-scale atmospheric circulation is surprisingly not a useful predictor for Tariff Days or Tariff Day runs. Local temperature, however, does show a ramp-up over one to three days previous to the onset of a Tariff Day run suggesting a very short lead time for the predictability of the onset of a Tariff Day run based on local temperature (see **figure 1-11**).

The following figures explore the predictability of the first Tariff Day in a run.

MSLP evolution leading up to the First Day of a Tariff Day run:

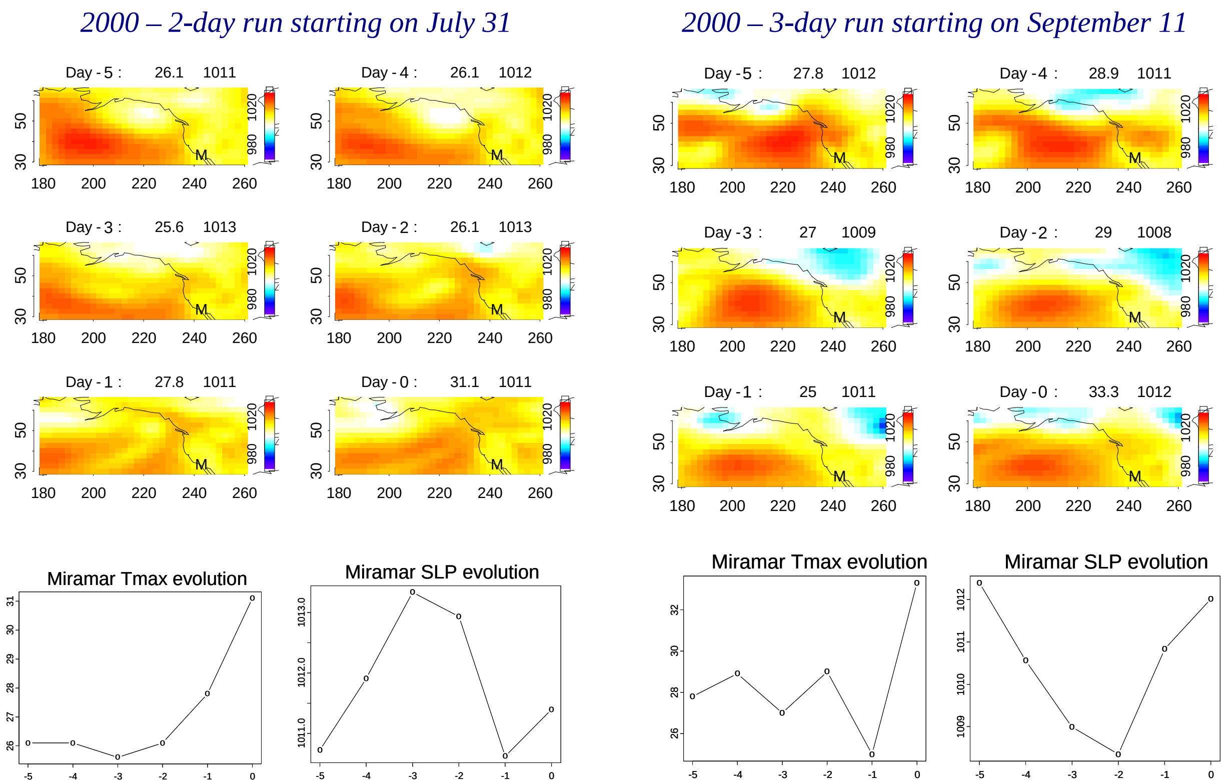


Figure 1-10. SLP Field, Local Temperature and SLP Evolution for July and September 2000

MSLP evolution leading up to the First Day of a Tariff Day run:

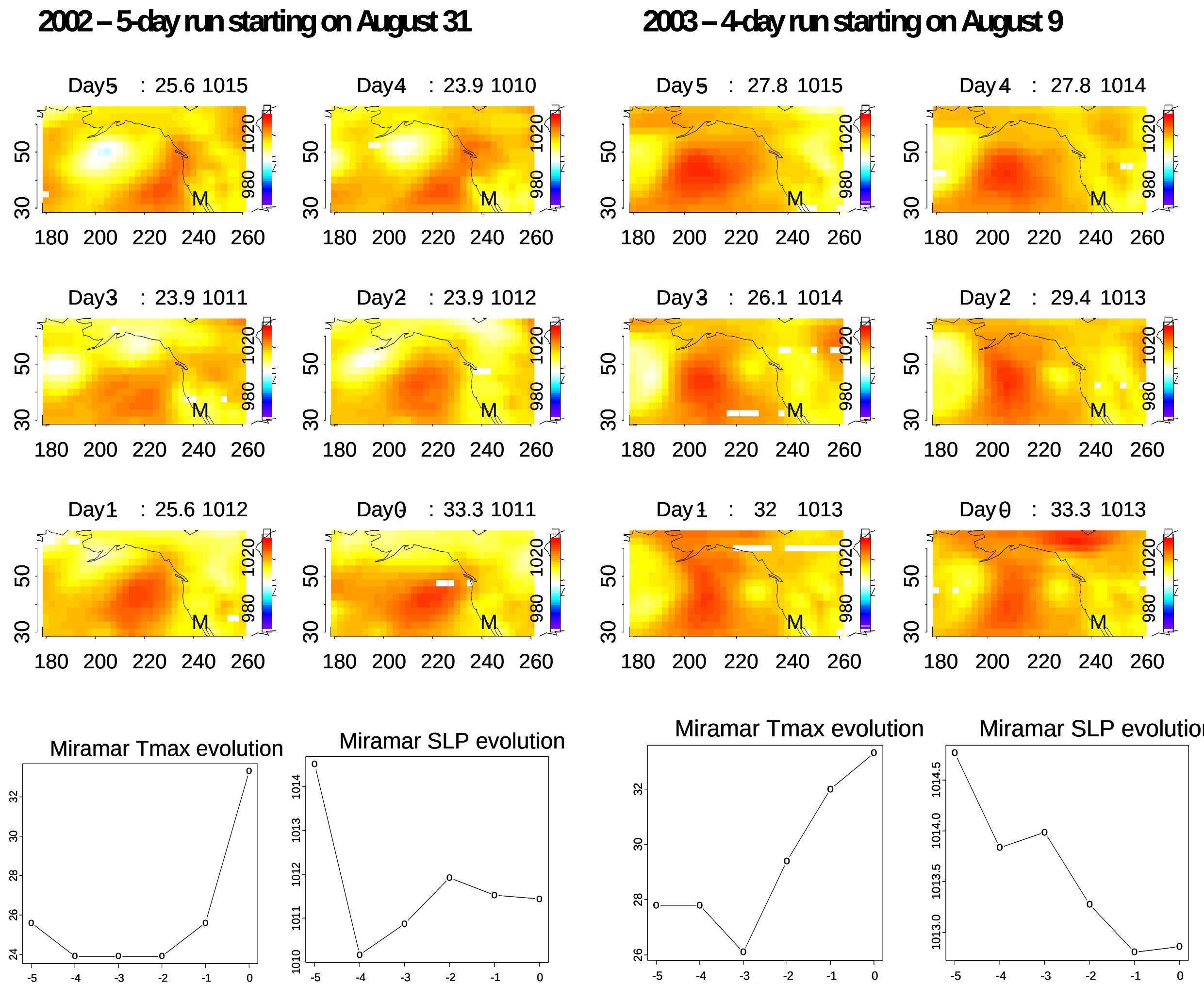


Figure 1-11. SLP Field, Local Temperature and SLP Evolution August 2002 and 2003

Reconstruction of JJASON mean of Tariff Day season intensity (mean Tmax) and timing (mean Julian day) at Miramar (red) and using SD Airport and Cuyamaca stations (black). Correlation of the reconstructed indices with March Sea Surface Temperature.

In **figure 1-12**, records at neighboring stations where longer records are available than at Miramar, were used to reconstruct average seasonal Tariff Day timing and intensity back to 1948, allowing for a seasonal analysis of the dependence of Tariff Day seasons on large-scale climatic conditions in the Pacific Basin. This figure shows that seasonal average intensity and timing of a Tariff Day season depend on large-scale climatic factors in the Pacific that are set in place months before the onset of the Tariff Day season. However, the relationship is weak, suggesting that seasonal predictability is limited.

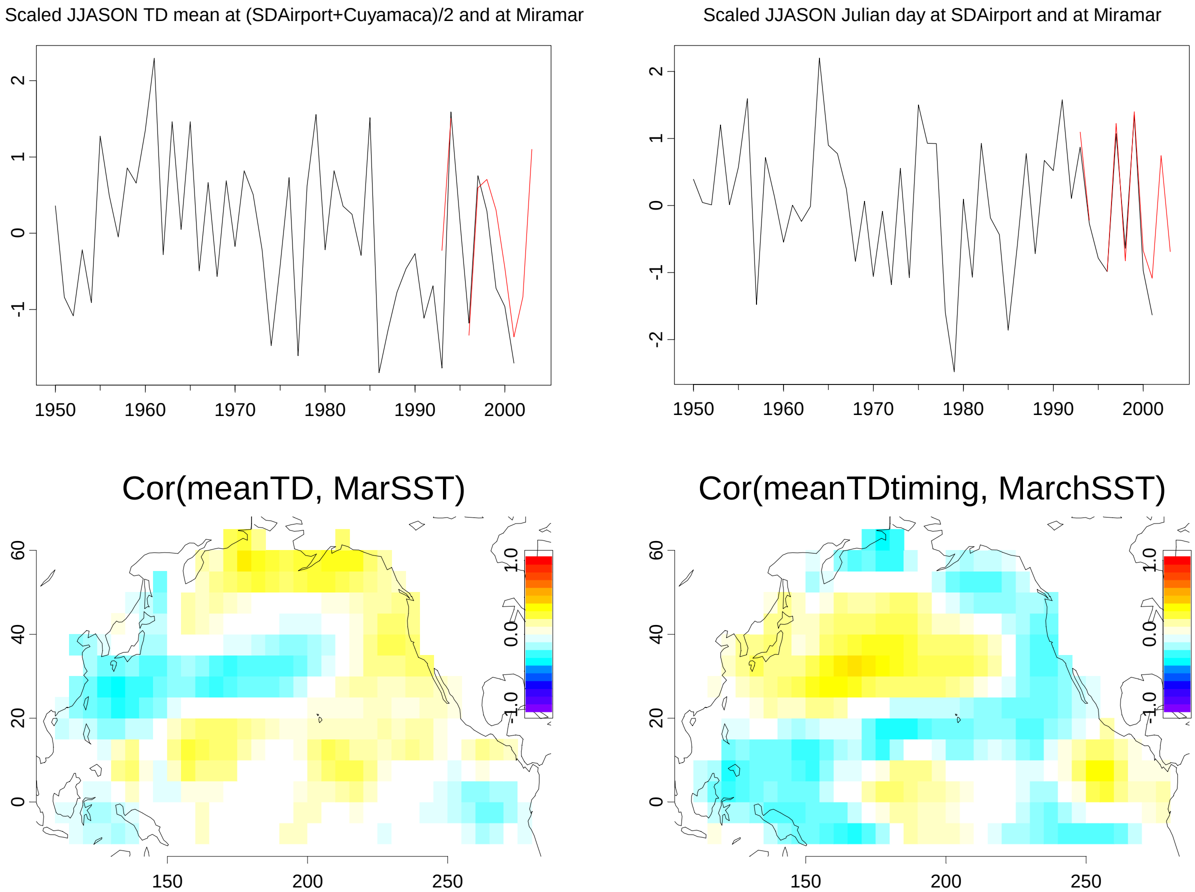


Figure 12. Reconstruction of Average Seasonal Tariff Day from 1948 to Present

Tariff days plotted together with Brawley – Lindberg Field (SAN) Tmax difference (i.e. inland-coastal Tmax, black curve) and SAN Tmax (green curve) with their respective means.

Figure 1-13 attempts to explore the synoptic relationship between coastal marine layer intrusions and Tariff Day occurrences. As expected, an association is found, however, the lead time is too short to be used for predictive purposes. Nevertheless, the association between marine layer intrusions and Tariff Day occurrences is strong and deserves more scrutiny, as the marine layer events may be climatically more predictable than Tariff Day onset. In this case, a predictive tool for Tariff Days may be possible. However, predicting marine layer intrusions is a rather complicated problem and is beyond the scope of this pilot study. Small Brawley – SAN Tmax difference indicates clear coastal conditions.

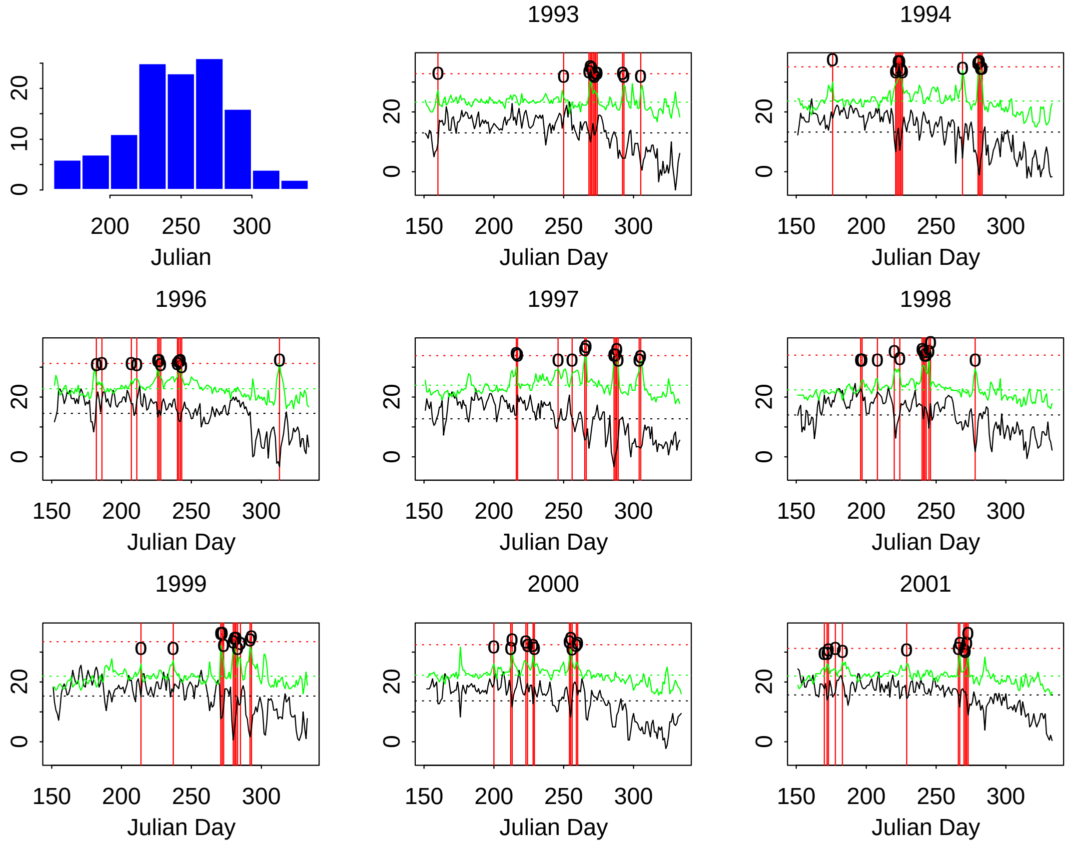


Figure 1-13. Marine Layer Influence

Estimating the Cost Effectiveness of Improving the Targeting of Load Management Tariff Periods

In a meeting with SDGE it was stated that the tariff event program had a goal of achieving a significant penetration in the market to clip 5% of SDGE system peak. The average system peak and off-season peak were presented in the earlier analysis of peak days with and without constraints. Average summer peak load is about X MW. With better targeting of loads to temperature sensitive time periods, an additional X MW's can be potentially obtained when considering (with some constraints) the 12 highest peak load days that were experienced in 2003 (see **figure 4-1**.). This means that if the goal of 5% of system peak were to be realized at the higher periods, an additional X MW's could be produced if the 5% goal were issued for the 12 highest peak load days using 2003 as a base. Experience in the summer 2001 electricity crisis showed that with aggressive marketing and a renewed sense of urgency, peak demand could be clipped by up to ten percent of peak load. Thus, the five percent target is more modest, compared to that which can be achieved in more urgent time periods with a call to action. However, the ten percent peak load reduction level through voluntary curtailments is not sustainable over time unless ongoing urgent appeals and corresponding pricing and other financial incentives are provided.

Furthermore, as the program is designed to be implemented in all seasons (in the pilot stage) many events are 'called' when the avoided costs to the power system could be negligible. While the current SDG&E assumed avoided cost is $85-kW-yr (the equivalent value of a peaking unit in Southern California), the actual avoided cost in non-summer peaking periods could be much lower. The avoided cost taking into account potential power plant spinning reserve during off peak periods might even be zero. In fact, the program in these cases may add cost.

Nevertheless, adhering to the current planning assumption of $85/kW-yr and recognizing that the goal is 5% of peak, the total potential value of the program is $X per year with the assumptions listed below.

Over a ten-year period, this program would amount to $X (in present value terms) – a benefit significantly higher than the current pilot program costs.

With better targeting the program to the higher peak periods, this would result in an approximately $X million per year, and a ten year present value of $X. The dollar savings appear to be significantly more impacted than the modest level of incremental demand that would occur.

As mentioned earlier the current program is only hitting about X% of the highest load periods and the improved forecasting approach could increase this to about a 6% hit rate, so the program is still capturing peak load reductions in periods when a substantial amount of spinning reserve and non-peaker unit generation is occurring.

(Costs rescinded)			

Key Assumptions

The power savings are an estimated 5% of an average summer peak load. The 5% was a stated goal of SDG&E for the program. The marginal generation capacity costs were also a California utility and SDG&E collaborative assumption given to SAIC/Scripps. This number was compared to the actual price and ten-year life cycle cost of a peaker for the California market and verified to be a reasonable cost assumption. Peaker costs outside of California, will be brought to market at substantially less cost. Annual savings calculations are shown in table 1-1. Inflation and discount rates are reasonable and consistent with industry assumptions for planning studies of this nature.

Cost estimate when well over 90% of the program occurs in periods that are either in the shoulder or off peak periods, is likely to overstate the benefits of the program at the pilot stage of performance and load impacts. The estimated savings shown in Table 1-1 are more indicative of potential benefits of the program if it were successful in shaving peak load during the highest peak load periods when peaking units were, in fact, operating or if spot power purchases occurred in the market that were either at $85/kw-yr or higher. It is quite likely that if load constraints and spot purchases were made to meet the load, chances are that the avoided cost could even be much higher than the $85/kw-yr value that is used.

5. Conclusions

The major conclusions from this analysis of the tariff demand response program are the following:

- Perfect knowledge of future loads would allow the selection of days with an average load 12-18% higher than usual on a summer afternoon.
- Constraints (2/week, 5/month) drop that to 11-15%. Weekly constraint is most important.
- Using temperatures at Miramar (as opposed to actual loads) introduce another level of uncertainty. With constraints, the best possible is 7-11%.
- Year 2003 was only X% of hitting the peak load. It may be possible to improve. Sources of possible improvement include statistical approaches, or using forecasts of weighted temperature rather than just Miramar. In addition, further improvement may be possible by changing the design of the program and attempting to focus the events during the periods with the highest historical probability of when peak loads would occur.

Figure 1-14 shows that the Mean Tmax evolution around an average single tariff day or the first day in a run of tariff days (day zero: black curve) and around any old tariff day (day zero: red curve). Tariff days are persistent. While the single and first days are not predictable (at least not by auto-regression) four days in advance, they may be predictable with 2-3 day lead-time. Tariff days other than first or single days are persistent and are predictable by auto-regression with a longer lead-time. This is the characteristic that allows a useful super-simple prediction scheme, as described above.

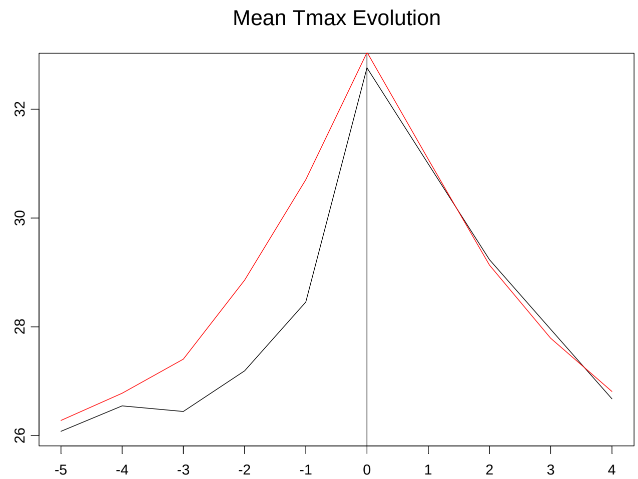


Figure 1-14. Prospects for Prediction Chart

The cost/benefit analysis of the pilot program uses an avoided cost based on the "peaker method". It has the drawback that it assumes forecasts achieve 100% skill in hitting peak periods. Despite that, electrical procurement cost savings are still obtained on days with higher than average load (costs) using the statistical forecasting scheme outlined here. This results in a direct savings of $X/yr. The final economic benefit of this program, then, is between the minimum attainable savings of $X/yr and the ultimate peaker method savings of $X/yr.

A more accurate reflection of program benefits are likely to be based on:

1. The actual hours when program events were called
2. Avoided capacity and congestion costs at the nodal points where the loads were reduced
3. The sum of all loads and avoided costs for each hour an event occurred. No such information was available and this information at a nodal or zone area might be too difficult to acquire unless the addresses of the participants were mapped to the zones in SDG&E's service territory. Given the proportionality of the loads occurring in off peak periods, the greater the likelihood that spinning reserves might be operating and avoided load and energy would be minimal.

This Page Intentionally Left Blank

Chapter 2: Use of Ensemble Forecast Models to Reduce Cal ISO Load Forecast Error

1. Introduction

The Cal ISO benchmark report of Deliverable 2 presented the details of the "model consensus" approach that Cal ISO uses to incorporate weather forecast information into the load forecasting process. This study establishes the value of using ensemble weather forecast probability distributions to significantly reduce the Cal ISO exposure to weather-related demand forecast errors. Given a collection of forecasts from different models, the traditional "model consensus" approach that Cal ISO uses simply averages the Model output to form a single forecast. In this report, ensembles of forecasts over different models and different initial conditions under the same model are considered. Rather than merely average this data, a variety of different methods for combining these simulations into a probability forecast are explored. The use of multiple models allows some accounting for model error, and it is found that the best performance does indeed come from a multiple model ensemble, called PF1 (Probability Forecast 1). This model, along with a probabilistic model build solely on AVN output, are explored further.

Employing probabilistic forecasts allows new approaches to risk management. Since the cost of the demand (temperature) exceeding the forecast can be much more than the cost of demand (temperature) falling below the forecast by the same amount, the best strategy is rarely (if ever) to go with the model consensus. A probability forecast provides many different levels, called isopleths. The most economically advantageous isopleth to use will depend on the cost function, but can easily be determined empirically. In this case, it is found to be about the 55% level for both PF1 and PF_AVN. This approach is, effectively, using the forecast distribution to automatically include a safety factor based on the difference in the relative penalties for over forecasts and under forecasts. This allows one to quantify weather related demand forecast confidence in a more quantitative and automatic manner at a relatively low cost.

This chapter contains details of the procedure used to select model simulations and the ensemble verification methods introduced to select PF1 from among the other probability forecast models considered. The impact of small data sets on the robustness of the statistical results (that is, whether the improvements of 2003 would be observed in 2004) is a major concern in any study of this kind. All initial model selections were done under "drop-one-out cross-validation" (rebuilding the model for the entire summer to forecast each day so that data on the current date is not included when forecasting a future day and the bootstrap significance levels are quoted in the tables.

Given the nature of the forecast usage and initial results on the information content in long range forecasts, this report focuses on short lead forecasts. The most extensive results are provided for a lead time of one day. The information in the forecasts decays rapidly, and the models available to form the ensemble changes as well since many of the models are only available for lead times beyond three days. It is shown that in the longer range there is little if any skill above climatology.

Finally, we note that this use of ensembles does not assume that the PF1 forecast is an objectively accurate probability forecast. If it were, then a different utility-based optimization would be in order. Rather we take the distribution as the best information available and extract the best demand forecast isopleth empirically.

2. Retrospective Weather Forecasting Results

From Simulations to Probability Forecasts: Overview

This section documents the probabilistic weather forecast analysis. Forecasting the region temperatures is reported and the transition from temperature probability forecasts to economic performance forecasts is discussed. The operational numerical weather prediction (NWP) station forecasts are translated into probability forecasts for daily maximum regional temperature (Tmax) in each of the four Cal ISO regions. The method then determines which isopleth of the forecast distribution maximizes expected economic utility. Current forecast accuracy reflects the use of summer 2003 data in both construction and evaluation with full (drop one out) cross-validation in the construction of the individual predictors.

Several distinct methods for constructing the probability forecasts are contrasted and described in detail below. The individual NWP forecast products used are given in Table 1 while a guide to the stations used is given in Appendix 1. The methodology and results are presented focusing on the Bay Region as an example. The details for all of the four regions are also presented.

Temperature as the Forecast Target

For each day of the summer of 2003, the target is the regional temperature as defined by Cal ISO. These targets is illustrated for the four regions in **Figures 2-1** to **2-4**. The aim is to produce a skillful probability forecast for these targets. In terms of forecasting weather variables, skill is measured relative to climatology. Given these results for the forecast weather variables, economic impact of the forecasts are evaluated in terms of dollars. For most NWP models in this study, the probability forecasts are formed by combining the output of an optimized linear predictor with its historical error statistics ("dressing" the forecast). For the NCEP forecasts, the initial condition ensemble itself provides information on the probability forecast on a given day. The marginal value of these various forecasts is then considered.

The operational value of the forecasts depends on the way in which they are interpreted. This study suggests that using various forecast isopleths is the best approach. Action firmly based on one of these isopleths rather than some "model consensus mean" will significantly decrease the user's weather risk exposure.

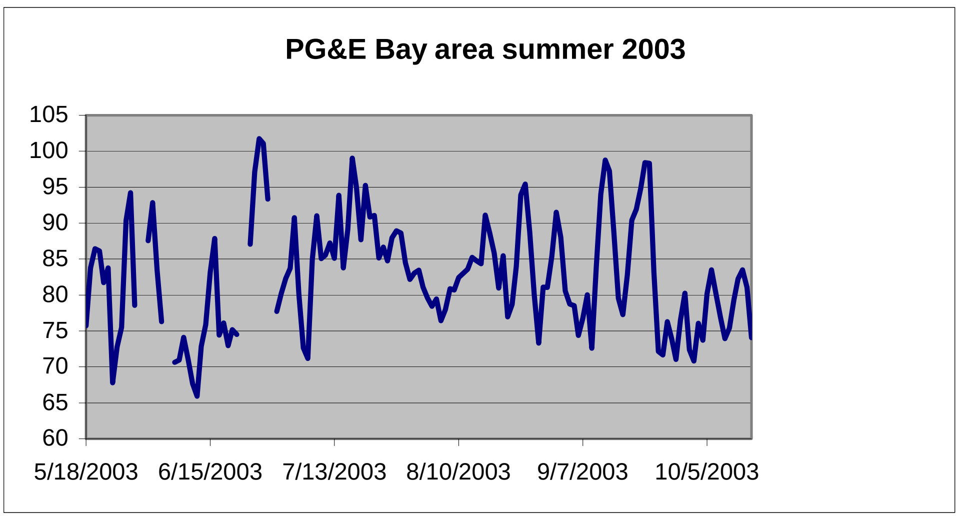


Figure 2-1. The Time Series of Cal ISO Bay Region Temperature Over the Summer 2003

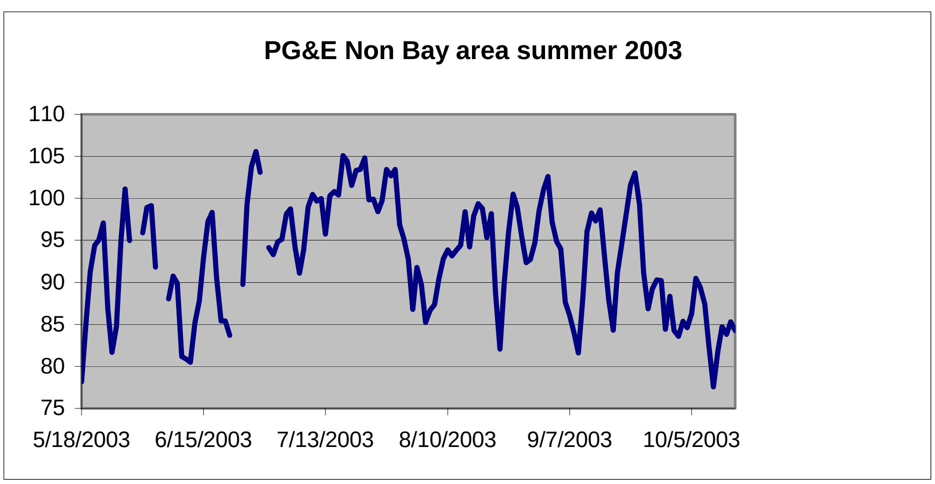


Figure 2-2. The Time Series of Cal ISO Non-Bay Region Temperature Over the Summer 2003

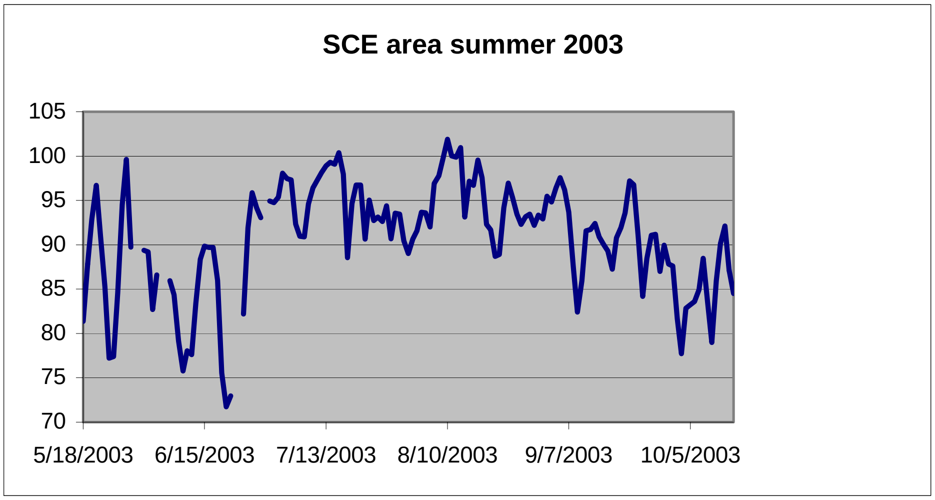


Figure 2-3. The Time Series of Cal ISO Region Temperature Over the Summer 2003.

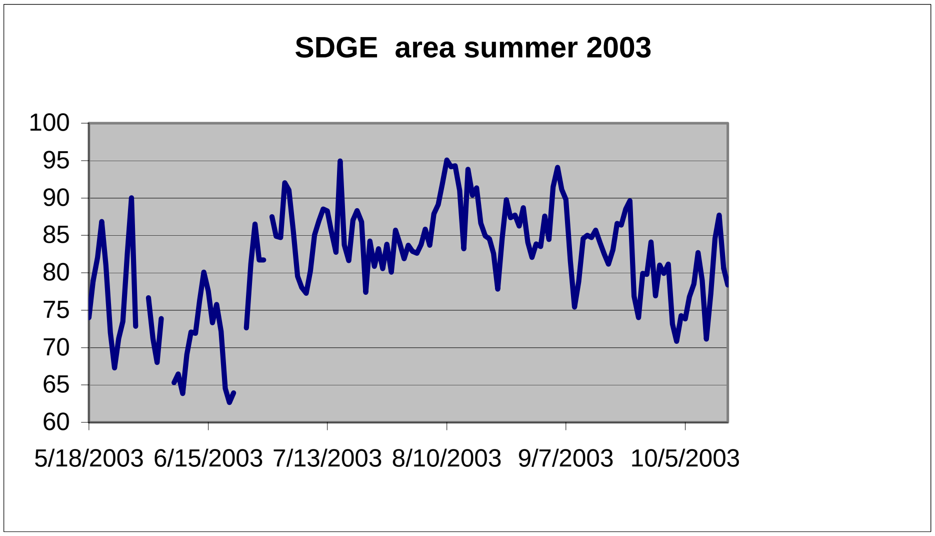


Figure 2-4. The Time Series of Cal ISO Region Temperature Over the Summer 2003.

The popular alternative to evaluating probability forecasts is to base decisions on the so-called “model consensus” forecast (see Fritsch et al, 2000, Weather and Forecasting, 15:571). Although this traditional approach aims for an “optimal blend” of forecasts, it increases the operating risk of any user with an asymmetric utility function, such as the one for CalISO as demonstrated in **figure 2-5**. Given that under-forecasting the maximum temperature by 3 degrees is likely to be much more costly than over forecasting by the same amount, the utility functions of interest to Cal ISO are clearly asymmetric. This implies that benefits of the model consensus approach currently used do not apply to the temperature forecasts as used in this application.

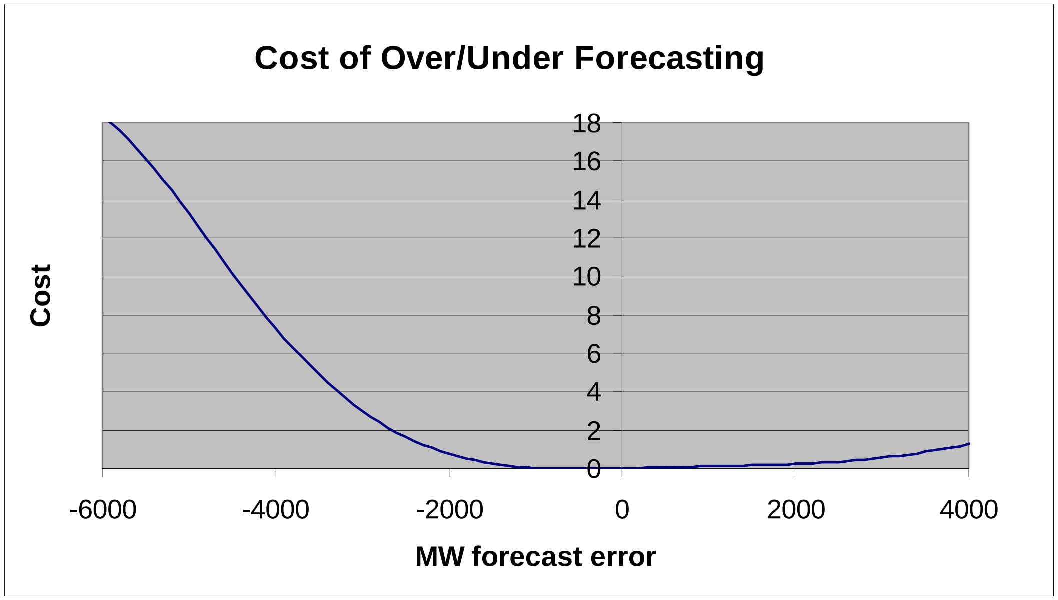


Figure 2-5. Cost Curve for Under/Over Forecasting From Cal ISO.
Note the asymmetry between over forecasts and under-forecast.

3. The Ensemble Approach: Selection and Verification

For each day, we also have a collection of predictors from various NWP models. We will consider this information in several distinct ways. The first is to combine information from each NWP model to form the best linear predictor for the region, and then combine these individual region predictions to form an overall best linear predictor. Typical forecast errors are then used to dress this point forecast with a Gaussian kernel. The kernel is simply a distribution, which replaces each point forecast (which can be thought of as a delta function). Kernels are commonly employed in density estimation, as an example in forecasting, the use of Gaussian kernels allows us to turn the prediction of 5 point forecasts (often treated as the sum of 5 delta functions) into the sum of 5 Gaussian distributions. The strength of this approach is that it provides a smooth forecast density, and allows us to treat the errors expected from having only a finite number of members in our ensemble. The word kernel is used to refer to the statistical distribution used to dress the point forecast, while calling the forecast itself the forecast distribution. The Gaussian kernel transforms the point forecast into a probability forecast (called PF1) Whereas PF1 uses the data from all the models simultaneously to produce a single linear point predictor, the second method constructs a distinct probability forecast based on the information from each model individually, where each has been weighted in terms of its skill *as a point forecast*. This probability forecast is called PF2. In the third method, the forecasts are formed as in method two, but they are dressed with different kernels that have been optimized to give a more skillful probability forecasts and referred to as PF3. Finally, the fourth and fifth methods exploit the fact that the NCEP model is run in ensemble mode and thus each forecast comes with an estimate of the "kernel of the day." So while the kernels used in the first three methods are static, that in method four and five allows the NCEP kernel to be derived from the ensemble itself and thus changes daily. At short lead times the dynamic kernel of the NCEP model is systematically too small, leading to very poor forecast scores. PF5 is a simple model that inflates the observed dynamical uncertainty.

Before considering tables of the results for various regions and PF models, an illustration of how a PF forecast can prove more useful in practice than the consensus forecast is presented. **Figures 2-6** and **2-7** show two different probability forecasts for the Bay Region. Both use model PF2 - one is from July 27, 2003 and the other from July 16, 2003. In the first, the consensus (that is the ensemble mean) is a fine estimate of the temperature; in the second, however, the mean is too low. PF1, however, has a wide probability forecast on this day, however, and the observed temperature corresponds to roughly the 70% isopleth of the probability forecast. Referring back to **figure 2-5**, the cost of a demand forecast based on a temperature over-forecast can be significantly less costly than a temperature under-forecast of the same amount – specifically the cost of a +4000 MW error is about one quarter of a –4000 MW error. Given this asymmetry in the cost curve of **figure 2-5**, there is significant utility in moving away from the mean (or median). Use of an isopleth other then the median (or the mean value) can avoid large losses. On one level this is nothing more than the common procedure of including a safety margin, but at a higher level a skillful probability forecast can not only provide a more objective standard for this "safety margin" but also allow the margin to adjust from day to day depending on the level of uncertainty in that day's forecast. . The key advantage over methods where the operator may add a safety factor, is that using an isopleth of the forecast distribution provides an optimized approach when the width of the forecast distribution is fixed, and in addition a safety margin that varies from day to day in step with variations in the reliability of the forecast.

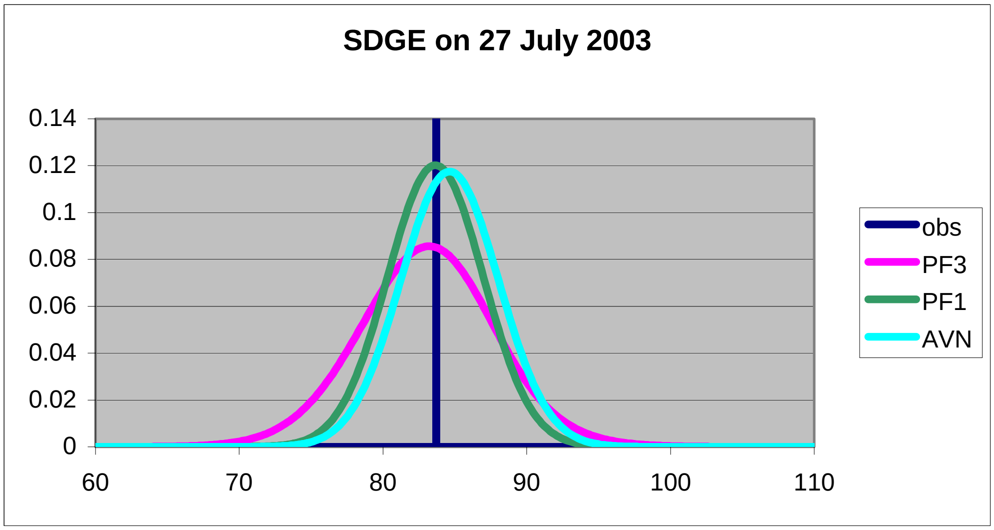


Figure 2-6. A "Good" Probability Forecast, Where the Verifying Observation Happens to Fall Near the Centre of the Forecast Distribution.
In this case PF1 out scores AVN and PF3, although all bound the verifying observation nicely.

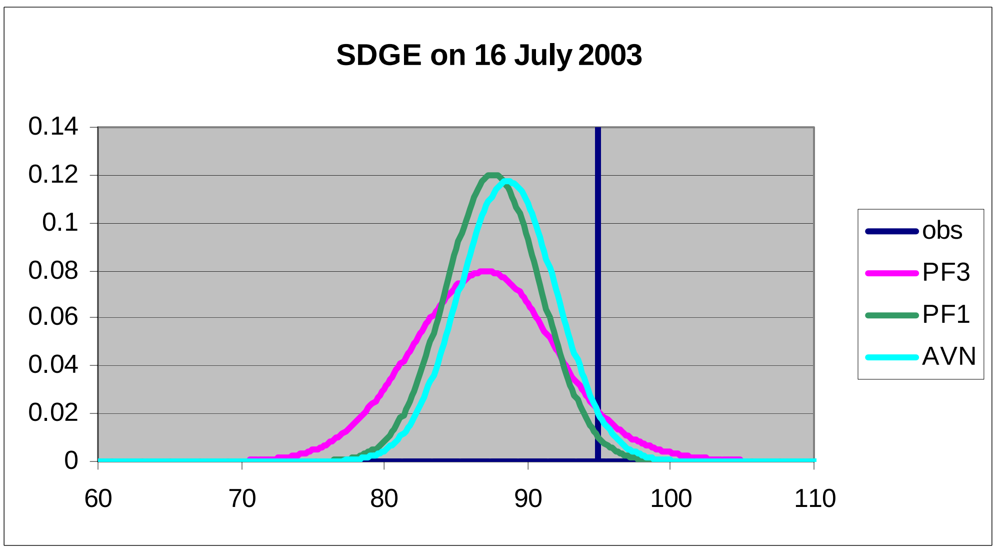


Figure 2-7. The Mean of the Forecast Distribution Shown Here is Significantly Below the Verifying Observation, But the Verifying Analysis Falls Well Within the Distribution of PF3.
Given that the penalty for under-forecasting demand is much greater than that for over-forecasting, using an isopleth well above the median can prove consistently profitable.

Three measures of forecast skill are used below. All point forecasts are evaluated in terms of the root-mean-square (RMS) error from the target. Probability forecasts are evaluated both in terms of their Brier Score (BS) and their Ignorance (IGN). The Brier Score is simply the square of the probability the forecast assigned to the target, averaged over all forecasts. Ignorance, on the other hand, is the negative of the log of the probability assigned to the target, again averaged over all forecasts. Thus while a small value of the RMS error or a small IGN score indicates a more skillful forecast, a larger value of the Brier Score indicates a more skillful forecast. The relationships between these measures of forecast skill are discussed in Roulston and Smith 2002 (Evaluating probabilistic forecasts using information theory, *Monthly Weather Review* 130 1653), Roulston and Smith, 2003 (Combining Dynamical and Statistical Ensembles, *Tellus* 55A, 16), and the reference therein.

Performance statistics for the PF forecasts for each of the four Cal ISO regions are given in the next section. The next step was to evaluate the performance of various isopleths in terms of economic value for each PF model. The contribution to performance was contrasted with an estimated cost of each component NWP data stream. A single PF model structure was selected for further study. The fact that the forecasts were interpreted in economic terms for the same summer in which they were constructed raises issues of cross validation. Ideally the 2003 data will also be evaluated using a model constructed from data covering the summer of 2002. Four PF models are constructed with drop-one-out cross-validation and then bootstrap uncertainty estimates are provided for each skill score in the tables below.

4. Relative Weights for Deriving Probability Forecasts

The four methods outlined above are now presented in detail. The key point of this section is that five distinct probability forecasts methods (PF1, PF2, PF3, PF4 and PF5) are constructed from the four models. The method of construction is fully specified in this section. An understanding of the remainder of the section, which contrasts the skill of these methods, does not require a full understanding of the construction of each PF.

Method One: The Combined Best Linear Predictor (PF1)

The first step was to construct a single probability forecast for a given region using the output from each weather model simulation (all the forecasts from a given model for the stations within the target region) At this time, Tmax models have been built based upon seven different sets of model simulations, namely AVN (including AVN-mos), ETA (including ETA-mos), MRF, NGM-mos, ENS and MEM. The last two (ENS and MEM) include ensemble information.[1] All historical forecast data was supplied by QuantumWeather, the only exception being Forecast information from Cal ISO which they used in their model. All our modeling results below are based on the QuantumWeather data. Model products are denoted by their QuantumWeather lables, the **table 2-1** below provides the standard product definition and a location for further information on that product (Appendix 2). Refer to the various input variables provided from each of these weather forecasts as a package, thus for each model and for each day there is a package of NWP forecast information and a verification target (the Cal ISO Regional Tmax).

Table 2-1. Inputs to the Seven Models, Time Scales, Variables for One-Day Ahead Forecasts

Package No	No	Model	Number of Predictors	Model run time 1	Delay 1	UTC time 1	Model run time 2	Delay 2	UTC time 2
1	1	avn	2	18	0	18	21	0	21
	2	avn_mos	2	18	0	18	21	0	21
2	3	avnext	0						
	4	ens	1	24	0	0			
	5	ens_max	1	24	0	0			
	6	ens_mean	1	24	0	0			

[1] Additional forecast products and models not in the draft report have now been included; the data sets for some forecast products were found to be corrupt and are excluded from the original analysis (for instance the ETA-mos). The main difference found from including the ETA-mos is seen in the ETA-new results, which consistently outperform the original (ETA) package.

	7	ens_min	1	24	0	0			
	8	ens_sdn	1	24	0	0			
	9	ens_sdp	1	24	0	0			
3	10	eta	2	18	0	18	21	0	21
	11	eta_mos	2	18	0	18	21	0	21
4	12	mem	1	24	0	0			
	13	mem_max	1	24	0	0			
	14	mem_mean	1	24	0	0			
	15	mem_min	1	24	0	0			
	16	mem_sdn	1	24	0	0			
	17	mem_sdp	1	24	0	0			
5	18	mrf	1	24	0	0			
	19	mrf_mos	1	24	0	0			
6	20	ngm_mos	2	18	0	18	21	0	21
7	21	ens_mean	1	24	0	0			
	22	ens_sdp	1	24	(difference 22 - 21 used only for prob. forecasts)				

For each package we perform a singular value decomposition. This provides the weights of the optimal linear predictor from that package to the target, as well as daily forecasts and residuals for each package. The results for the four Regions are shown in **tables 2-2**, **2-3**, **2-4**, and **2-5**.

Table 2-2. Relative Performance Over the Summer of 2003 of the Individual NWP Packages, Shown for Bay Area.
Note that the AVN package and PF1 tend to have the best ignorance and Brier scores.

PG&E Bay Area Summer 2003							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.01	3.42	-1.15	0.09	7.68	0.43
2	ENS	0.00	6.35	-0.26	0.10	2.34	0.12
3	ETA new	0.01	4.06	-0.90	0.10	5.55	0.32
4	MEM	0.00	6.77	-0.16	0.15	2.23	0.10
5	MRF	0.00	5.77	-0.40	0.11	2.85	0.16
6	NGMMOS	0.01	4.23	-0.84	0.10	5.28	0.28
7	ENS new	0.00	5.83	-0.38	0.09	2.75	0.14
8	PF1	0.01	3.56	-1.09	0.09	7.11	0.43
9	PF2	-	-	-0.78	0.05	3.61	0.15
10	PF3	-	-	-0.86	0.04	4.05	0.13
11	PF5	0.00	5.83	-0.35	0.11	3.26	0.28
12	Cal ISO	1.31	4.31	-	-	-	-

Table 2-3. Relative Performance Over the Summer of 2003 of the Individual NWP Packages, Shown for Non-Bay Area.
Note that this time PF1 scores best, but again the AVN package and PF1 tend to have good ignorance and Brier scores, however ETA-new does well in this region also. Further tuning for PF3 could be of significant value of sufficient data is available to validate the tuning.

PG&E Non Bay Area Summer 2003							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.04	1.66	-2.02	0.11	35.54	1.83
2	ENS	0.01	3.83	-0.82	0.11	6.39	0.32
3	ETA new	0.04	1.51	-2.16	0.08	41.22	1.68
4	MEM	0.01	4.00	-0.76	0.08	5.91	0.28
5	MRF	0.01	3.52	-0.94	0.08	7.27	0.39
6	NGMMOS	0.02	2.13	-1.67	0.09	20.04	1.30
7	ENS new	0.01	3.41	-0.98	0.10	8.08	0.40
8	PF1	0.05	1.39	-2.28	0.09	48.17	2.36
9	PF2	-	-	-1.62	0.04	13.99	0.53
10	PF3	-	-	-1.83	0.04	18.96	0.56
11	PF5	0.01	3.41	-1.01	0.10	8.91	0.57
12	Cal ISO	-0.37	2.44	-	-	-	-

San Diego Edison Summer 2003							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.02	2.10	-1.64	0.16	22.30	1.22
2	ENS	0.00	4.30	-0.61	0.08	4.82	0.25
3	ETA new	0.01	2.86	-1.20	0.11	11.88	0.60
4	MEM	0.00	4.65	-0.50	0.11	4.56	0.20
5	MRF	0.01	3.71	-0.82	0.08	6.70	0.33
6	NGMMOS	0.02	2.50	-1.39	0.10	15.44	0.72
7	ENS new	0.01	3.92	-0.74	0.11	6.11	0.39
8	PF1	0.02	2.04	-1.69	0.11	23.53	1.29
9	PF2	-	-	-1.18	0.07	8.83	0.45
10	PF3	-	-	-1.30	0.07	10.42	0.50
11	PF5	0.01	3.92	-0.51	0.18	8.68	1.03
12	Cal ISO	0.33	2.59	-	-	-	-

Table 2-4. Relative Performance Over the Summer of 2003 of the Individual NWP Packages, Shown for San Diego Edison Area.

PF1 scores best, although AVN is nearly within the one sigma the bootstrap limits.

San Diego Gas and Electric Summer 2003							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.01	3.39	-1.11	0.11	8.16	0.56
2	ENS	0.00	5.22	-0.49	0.08	3.18	0.19
3	ETA new	0.01	3.82	-0.94	0.12	6.78	0.31
4	MEM	0.00	5.43	-0.43	0.19	3.41	0.16
5	MRF	0.00	4.52	-0.70	0.07	4.36	0.22
6	NGMMOS	0.01	4.19	-0.80	0.08	5.44	0.31
7	ENS new	0.00	4.98	-0.56	0.11	3.65	0.20
8	PF1	0.01	3.32	-1.14	0.10	8.63	0.48
9	PF2	-	-	-0.88	0.07	4.56	0.25
10	PF3	-	-	-0.92	0.05	4.82	0.20
11	PF5	0.00	4.98	-0.40	0.13	4.96	0.50
12	Cal ISO	2.11	3.88	-	-	-	-

Table 2-5. Relative Performance Over the Summer of 2003 of the Individual NWP Packages, Shown for San Diego Gas and Electric Area.

Note that the AVN package and PF1 again have overlapping bootstrap ranges, but PF1 is again scoring better

Note that the PF model called AVN_Package uses different inputs and is fit differently than the Cal ISO use of AVN model output. This explains why AVN_Package performs significantly better than the Cal ISO AVN, although both are using weather forecast information from the AVN model.

The next step combines these individual package forecasts into a single forecast. Method One does this by finding an optimal linear combination. Method Two uses an alternative approach.

Given the six forecasts from the packages, Method One now forms relative weights for the optimal linear predictor combining these inputs, again using a singular value decomposition where the input matrix is now composed of the output from the optimal linear predictors of the individual NWP packages. The target is the observed Tmax. The relative weights are given in **table 2-6**.

Table 2-6. Sample Relative Weights per Package and Performance.

This example is for PF1 forecasting KSFO; a similar set of weights is computed for each model for each station in each region. The resulting table is somewhat unwieldy. Average weights will be included the Supplemental Material (in soft form).

Station	AVN	ENS	ETA	MEM	MRF	NGM MOS	CONST	Bias	Standard Deviation	Mean Ignorance	Brier Score (*10^3)
KSFO	0.33	-0.10	-0.17	0.21	0.15	0.58	0.00	0.10	4.35	4.17	4.86

Method One now constructs a probability forecast defined by a Gaussian distribution centered on the forecast value with a standard deviation defined by the historical error statistics. The standard deviation does not change from day to day due to the synoptic state, but could be adjusted with month (using 2002 data from 2003, or 2003 data from 2004, and so on), or computed over a sliding window (to account for changes in the model(s)).

Method Two: Union of Distributions From Linear Forecasts (PF2)

From **table 2-2**, it was demonstrated that the individual forecast from each package has variations in skill. In a non-risk neutral setting, it appears that one model provides (statistically useful) early warning of a large change. In terms of a linear predictor, a model will be penalized for doing this, and given a lower weight relative to a more conservative model which yields a better root-mean-square error forecast but under-forecasts every single large change.

The approach used in Method Two allows the possibility that such information can be exploited by combining the forecast distributions of the individual packages, rather than placing a distribution about their combined forecasts. This naturally allows a wider distribution on days when the individual packages disagree.

The simplest implementation of this approach (PF2) is to sum a series of Gaussian distributions, one from each model, where the models are equally weighted and the width of each distribution is simply the standard deviation of that package's historical forecast errors. This defines PF2.

Method Three: Kernels Optimized with Information Theory (PF3)

In the previous approach, the kernel width was based upon the historical skill of a package at delivering point forecasts. In PF3, a more relevant information theoretic skill is optimized. Again the forecast of each package is dressed with a Gaussian kernel, but in PF3 both the width of that kernel and the weight given to each model is adjusted in terms of overall model performance.

Optimal construction of PF3 would involve a high dimensional optimization problem. There is not sufficient data to justify this approach. Rather, a series of PF3 models was built, adjusting the weight of each model, one model at a time. This approach has the added advantage that the cost and reliability of delivery of each model can be factored into the evaluation.

Another advantage of PF3 (and to a lesser extent PF2), is variation in the width of the PF3 probability forecast from day to day; several predicted probability density functions are shown in **figure 2-8**, along with the verification. In this case the verifying observations are below the mean of PF3 distribution: employing an isopleth above the mean would not help in this case. Nevertheless, this distribution is visibly wider than typical, and hence there was some prior warning that this was a particularly uncertain forecast. The extent to which this knowledge could be acted upon depends, on the detailed options available to Cal ISO. Such early warning can be found in a number of the summer's forecasts.

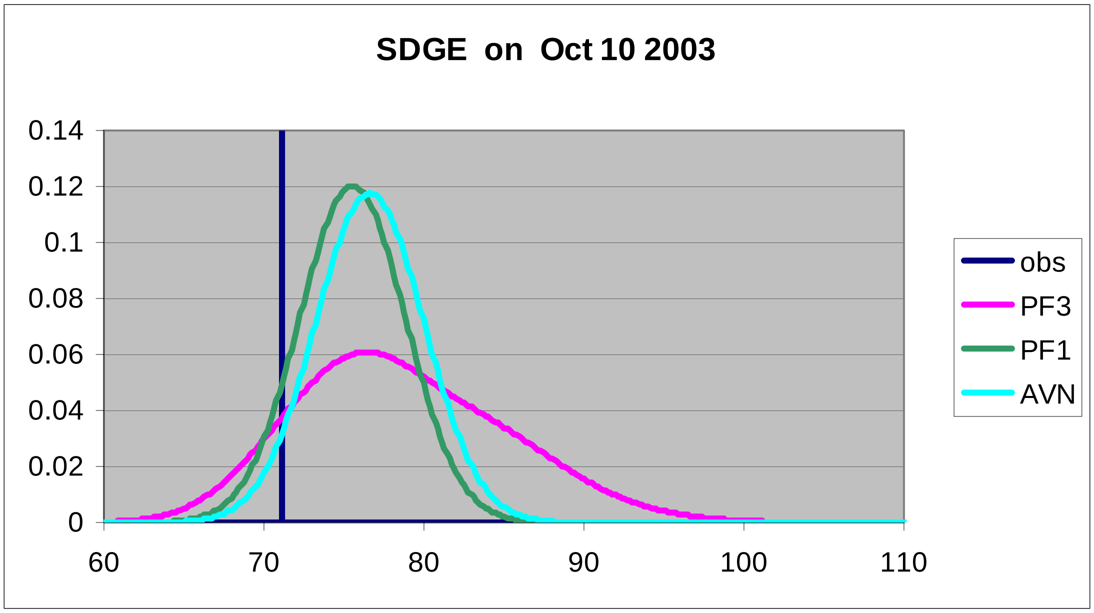


Figure 2-8. Illustration of a Relatively Wide Forecast From PF3 on Oct 10, 2003 for SDGE; in this Case the Verifying Observation Fell Below the Median.

The advantage of the PF3 forecast here is that it was known to be uncertain when the initial decisions were taken and alternative strategies to cope with this 'irreducible uncertainty' could have been put into place for this particular day.

Method Four: Dynamically Determined Kernels (PF4)

This method allows the spread associated with those models that run an ensemble of initial conditions to change from day to day. For models which do not have ensembles, the components of predictor consist of the sum of Gaussians from the individual packages just as in PF2. Those models which do have ensemble information provide additional components centered on their forecast, but with a standard deviation which varies with the width of that NWP ensemble forecast. First the relative width and then the relative weight of the model are adjusted to improve the ignorance score. Although the value of the information in the ensemble spread is expected to increase with lead-time, no significant value in the use of the dynamical spread "as is" at any of the lead times tested was detected. As a result, no results for PF4 are included, rather, PF5 is introduced, where the kernel is based upon the dynamical kernel-of-the-day, but adjusted statistically to improve the skill scores.

Table 2-7. The First Line of the Table Shows the Average of the Standard Deviation of the Ensemble Members Over All Summer 2003 Days, for Each Region.

The lower eight entries show the inflation factor (for each region) required so that the ensemble spread of the dynamical ensemble will match, on average, equal that of the historical error distribution of that model or package. These results are not out-of-sample.

	PG&E Bay	PG&E Non Bay	SDE	SDGE
Ensemble Standard deviation	2.11	3.31	2.52	2.95
PF1	1.69	1.13	0.81	1.13
AVN	1.62	1.15	0.83	1.15
ENS	3.01	1.77	1.71	1.77
ETA new	1.93	1.29	1.13	1.29
MEM	3.22	1.84	1.84	1.84
MRF	2.74	1.53	1.47	1.53
NGMMOS	2.01	1.42	0.99	1.42
ENS new	2.77	1.69	1.56	1.69

Method Five: Dynamically Based Kernel Widths (PF5)

Given that the dynamically based kernel provided by the ensemble was systematically too narrow, a simple method of inflating the dynamical width by a fixed factor was used to increase the skill of the NCEP forecast. The inflation values are given in **table 2-7** below. Better methods for adjusting the ensemble based kernel width, for instance setting a lower bound based on the historical error in the mean. This would limit the utility of the forecast in identifying days with higher than normal predictability. For this application, however, there was no method which yielded significant skill to this dynamical-based kernel.

Comparison of Forecast Performance

The forecasts of these models were then contrasted, noting that PF1 is a Gaussian with a fixed width, that PF2 is a sum of (fixed width) Gaussian distributions and thus the width of PF2 is expected to vary from day to day. PF3 is again a sum of a different set of fixed width Gaussians. PF4 and PF5 have widths determined from a dynamical ensemble forecast.

The forecast skill for various methods discussed above is contrasted in **tables 2-2, 2-3, 2-4**, and **2-5**.

We see that PF1 and the AVN package consistently have the best scores. The bootstrap significance levels show that the performance of several of the methods is similar, but the fact that PF1 is consistently better than the AVN package outside the Bay Area suggests true skill. Tests on a second summer season will clarify this issue.

Beyond the issue of pure skill, there are other operational issues to consider. For instance, PF1 requires the entire NWP forecast dataset to be available every day, whereas PF2 can easily be constructed and interpreted even when one of the NWP packages is missing. While PF3 need not require all the NWP packages in the first place, it is unlikely to be robust if a NWP package that it ranks highly is missing on a given day. It is recommended that alternative PF models are developed and calibrated for use in the event that any one NWP information package is unavailable on a given day.

Interpreting Isopleths Other than the Mean

Consider how the observed target value is likely to appear when plotted over the forecast isopleths of a PF forecast. The target values will sometimes fall above the 90% isopleths. This is, of course, expected even for a perfect probability forecast, where about 10% of the targets *should* fall above the 90% isopleth, just as 50% are expected to fall above the median.

Forecast users with asymmetric risk need not base decisions upon the forecast mean value, even with a perfect probability forecast. Instead, they can plan based upon some other isopleth. In the case of Cal ISO where under-forecasts are more costly than over-forecasts, one is most likely to take an isopleth between 50% and 90%. The particular value chosen will depend upon a mixture of the relative costs of reality falling above or below the threshold chosen, the volatility of the quantity measured, and the quality of the forecast model in hand.

Temperature Forecasts Beyond Day One

The collection of forecast models above was developed for one day ahead forecasting. There is clearly a need for longer range probability forecasts. Lead times of 2 days, 3 days, 10 days and 14 days are explicitly considered. For days 2 and 3, our AVN package and PF1 both show significant skill above that of climatology These results are presented in the next section where, like the day one forecasts, they are translated into demand forecasts. Many of the numerical weather model forecasts that go into PF1 are only available for lead times less than 10 days, thus a 10 or 14 day PF package can not include the MOS forecast products or any ETA products. In addition the number of forecast reporting times also decrease with lead time. None of the PF packages demonstrated skill above climatology at day 10 or day 14. It is often observed that, at lead times of 10 days and longer, the ensemble tends to be too narrow. Including historical information on forecast accuracy, however, tends to yield a forecast distribution much too wide, sometimes extending beyond the range of historical observed temperatures (that is, outside climatology). Translating these extreme temperature distributions into demand forecasts yields fallacious results and they are not pursued further here. It would be interesting, however, to develop a package which combined these longer range forecasts with climatological data conditioned on the current observations. Of course, it may be that there is simply no information relevant to demand that can be extracted from these forecasts at these lead times.

5. Economic Impact Results

Objective

This section outlines the methodology for exploiting the information content of improved probabilistic weather forecasts in Cal ISO's demand forecasting and quantifying the likely economic impact. Specifically, we show how the probability distribution temperature forecasts for each of the four Cal ISO regions can be exploited to improve the expected economic performance of Cal ISO demand scheduling. This process consists of two steps, the first is to realize that the traditional approach of forming a single best-first-guess (or consensus) forecast is sub-optimal. Better economic decisions are almost certainly possible by *not* basing demand decisions on the best informed RMSE-minimum estimate of temperature implied demand. Rather, due to the asymmetric cost of over-forecasting and under-forecasting, expected economic benefits are maximized by planning according to a temperature value offset from the median forecast value, where the offset will vary with the reliability of the forecast. The second step consists of determining exactly which probability contour (the median? The 60%? The 75%?) or isopleth to use in a given situation. In short, the 55% isopleth of our probability model PF1 yields the most significant reduction of weather related costs.

As stated earlier, but worth reiterating here, operators currently "unsystematically" add a safety factor to the forecast to take into account the effect of the non-systematic cost function with which they are clearly familiar. Operators realize that there is a greater negative consequence for a large underforecast miss than for a long string of small overforecast misses, even when the latter costs more over the year. The ensemble methods presented here give them a "systematic" way to "improve their forecast.

In addition, the median of PF1 (or the AVN _Package) differs from the Cal ISO calculation of the demand forecast developed from their AVN station data. The method presented here demonstrates greater accuracy as well. It processes the AVN output in a better way and provides forecast distributions to allow a systematic optimization of the isopleths.

These results are robust under bootstrap resampling of the 2003 data, however the effect of using the data from a previous year, as required in practice, cannot be accounted for in this way. To evaluate the effect of having data only from previous summers, data from the summer of 2002 are considered.

From Temperature Probability Forecasts to Demand Allocation: Overview

For each day of the summer of 2003, the target was to minimize the implied weather-component of demand error, subject to the constraint that the total economic loss due to weather-related demand forecast error is minimized. The constraint is critical, given the asymmetric nature of the cost curve, which was illustrated in f**igure 2-5** above. Note that the cost of a 4000 MW over-forecast (that is, a temperature forecast above the temperature observed) is significantly less than the cost of a 4000 MW under-forecast. This implies that even if error distribution in the temperature-dependent demand was constant and symmetric about the mean every day (which it is not), then overall cost could be lowered by using a demand forecast based on a forecast isopleth greater than the forecast mean value, possibly the 60% contour (or if the distributions were all Gaussian, then perhaps the median plus a quarter of a standard deviation). The reason for this is obvious even if the implementation is not. By regularly paying a low cost, one mitigates a number of what would have been very large losses. The next section discusses how this is done.

Performance with Isopleth: An Example

What follows is an illustrative example of the use of isopleths. **Figure 2-9** shows the distribution of the MV forecast errors based on the sample of 147 days. The magenta curve reflects a Gaussian distribution with same mean and standard deviation as the observed distribution - namely, the mean and standard deviation in the summer 2003 season equal -131 (MW) and 1045 (MW). The lower tail is dominated by the two under-forecasts on June 16 and 17. The Gaussian is not a very good fit. The sample CDF appears steeper than the normal.

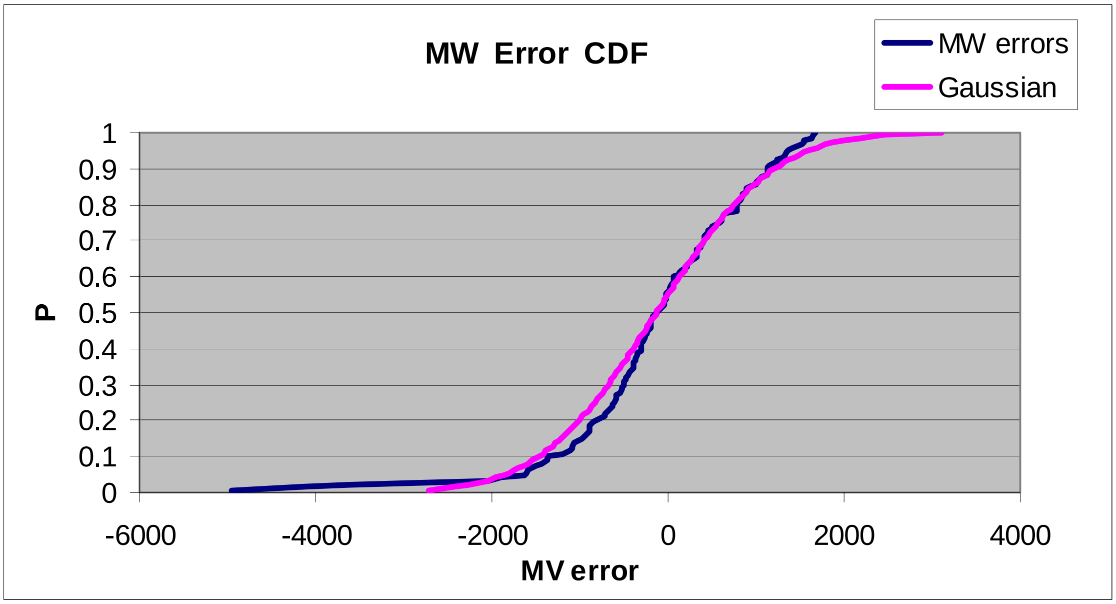


Figure 2-9. Cumulative Distribution Function for Megawatt Error

To illustrate the method, imagine if the operator had foreknowledge of the likely Megawatt error distribution (probabilistic temperature forecasts give that, as argued below the relationship between temperature error and demand error is roughly linear). If the fitted Gaussian distribution was resampled and the expected loss computed, the outcome is about 0.135 (differing slightly from the observed loss because the distribution is not, in fact, Gaussian). Since the distribution is known, however, one can move out on the 'over-forecast' side and observe a drop in the expected loss, as shown in **figure 2-11**. In this case, the expected minimum loss occurs at over half a standard deviation.

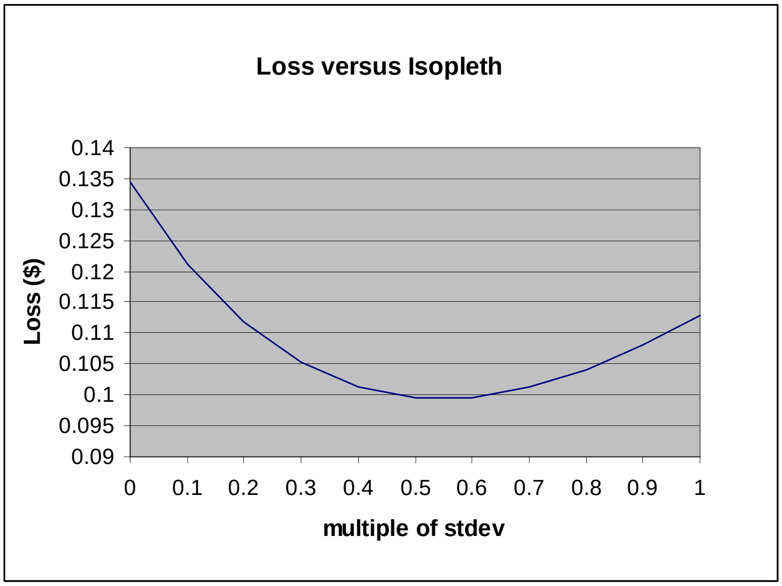


Figure 2-10. Expected Loss as a Function of the Isopleth, Measured in Terms of the Standard Deviation

Note that **figure 2-11** shows the MV forecast error versus the four-region temperature forecast error. (The four-region temperature forecast error is the sum of the ((signed)) forecast errors in the four regions.) Both the MW errors and the four-region temperature errors are defined as the forecast value minus the observed value, so a positive value means an over-forecast. Large positive values often coincide with all 4 regions being over-forecast. Comparing the graph with the table below, the two biggest under-forecasts on 16 and 17 June appear not to be obviously temperature related.

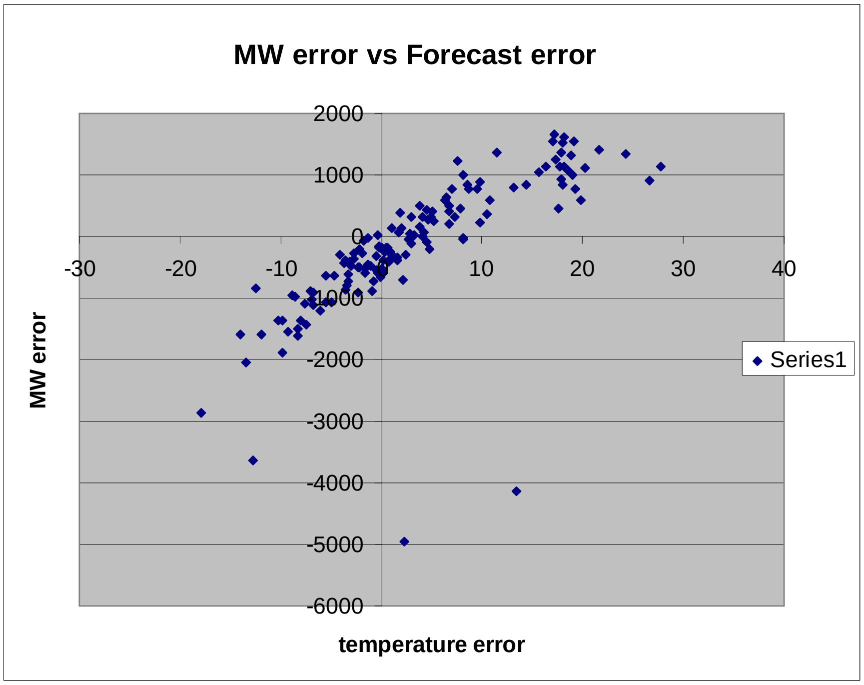


Figure 2-11. Scatter Plot of MW Error Versus the Sum of the Four (Signed) Cal ISO AVN Regional Temperature Forecast Errors.

The MW errors per day were taken from "Regression Forecaster vs. Observed MW (1).csv" while each regional temperature error was taken from "AVN ETA MRF day-ahead forecasts vs. observed[1].csv"

The five dates with the largest under-forecasts in the summer 2003 season can be identified on the **figure 2-11**. The corresponding dates are:

Date	Under Forecast
16-Jun-03	-4949
17-Jun-03	-4142
27-Jun-03	-3626
28-May-03	-2864
21-May-03	-2044

Regional Temperature Forecast Results

In this section new temperature forecasts for each of the regions are calculated. These results are fully cross validated (using drop-one-out estimation for each day, given the small data set available). These are the ignorance results and error metrics of our six packages, plus our PF1 and PF2 multi-model probability forecasts and the Cal ISO AVN as a reference. The major differences to the previous results are:

- Ignorance is measured against climatology, where climatology is determined from the best fit Gaussian to the observed distribution. The climatological distribution is assigned an ignorance score of zero (no skill), so the forecast values are now negative numbers, the smaller the ignorance, the better the score.
- The region temperatures are forecasts directly, based on the Cal ISO values provided. The available numbers of days in a season are restricted to only those days in which there are both the Cal ISO region observations and the relevant forecasts. In summer 2003, in Cal ISO data consists of a total of 144 days for which observations are available for all four regions. Due to unavailable forecasts, the total number of days available in the summer season between May 18 and October 15 differ in the

different regions: PG&E Bay area (116 days), PG&E Non Bay area (115 days), San Diego Edison (116 days), San Diego Gas & Electric (115 days).

- Comparisons of forecast performance in the regions between these new retrospective weather forecasts and the existing Cal ISO AVN forecasts have been based on the identical set of days in the summer 2003 season.
- The package ETA now also contains the ETA MOS point forecast. None of the retrospective package has a true model Tmax as predictor. (that is, at this time each value input as a model predictor corresponds to a model hour temperature value.)

PG&E Bay Area					
Number	Package	Bias	Standard deviation	Mean rel. Ignorance	Brier Score (*1000)
1	AVN	0.03	3.40	-1.15	7.68
2	ENS	-0.03	6.32	-0.26	2.34
3	ETA	-0.04	4.04	-0.90	5.55
4	MEM	0.26	6.75	-0.16	2.23
5	MRF	0.05	5.74	-0.40	2.85
6	NGMMOS	0.04	4.22	-0.84	5.28
7	Cal ISO AVN	1.31	4.31	-	-
8	PF1	0.02	3.55	-1.09	7.11
9	PF2	-	-	-0.86	4.02

PG&E Non Bay Area					
Number	Package	Bias	Standard deviation	Mean rel. Ignorance	Brier Score (*1000)
1	AVN	0.01	1.66	-2.02	35.54
2	ENS	-0.02	3.83	-0.82	6.39
3	ETA	-0.02	1.51	-2.16	41.22
4	MEM	-0.05	4.00	-0.76	5.91
5	MRF	0.00	3.52	-0.94	7.27
6	NGMMOS	0.02	2.13	-1.67	20.04
7	Cal ISO AVN	-0.37	2.44	-	
8	PF1	-0.01	1.39	-2.28	48.17
9	PF2	-	-	-1.73	17.02

San Diego Edison					
Number	Package	Bias	Standard deviation	Mean rel. Ignorance	Brier Score (*1000)
1	AVN	-0.02	2.10	-1.64	22.30
2	ENS	0.06	4.30	-0.61	4.82
3	ETA	-0.06	2.86	-1.20	11.88
4	MEM	-0.10	4.65	-0.50	4.56
5	MRF	0.01	3.71	-0.82	6.70
6	NGMMOS	0.02	2.50	-1.39	15.44
7	CalIso AVN	0.33	2.59	-	-
8	PF1	-0.02	2.04	-1.69	23.53
9	PF2	-	-	-1.28	10.21

San Diego Gas & Electric					
Number	Package	Bias	Standard deviation	Mean rel. Ignorance	Brier Score (*1000)
1	AVN	-0.01	3.39	-1.11	8.16
2	ENS	0.02	5.22	-0.49	3.18
3	ETA	0.03	3.82	-0.94	6.78

4	MEM	-0.19	5.43	-0.43	3.41
5	MRF	0.03	4.52	-0.70	4.36
6	NGMMOS	0.01	4.19	-0.80	5.44
7	CalIso AVN	2.11	3.88	-	-
8	PF1	0.01	3.32	-1.14	8.63
9	PF2	-	-	-0.94	5.07

Note that the direct retrospective region forecasts show the same systematic differences in predictability in these 4 regions as seen in the Cal ISO AVN forecast. The Bay area is the most difficult to forecast in each case. Only in the Bay area does the forecast PF1 not show the best results with respect to three forecast metrics standard deviation of error, mean ignorance and Brier Score. Note also that the AVN Package and PF1 Package tend to outperform the pure Cal ISO AVN forecast.

Temperature Isopleth for Bay Area PF1 Forecasts

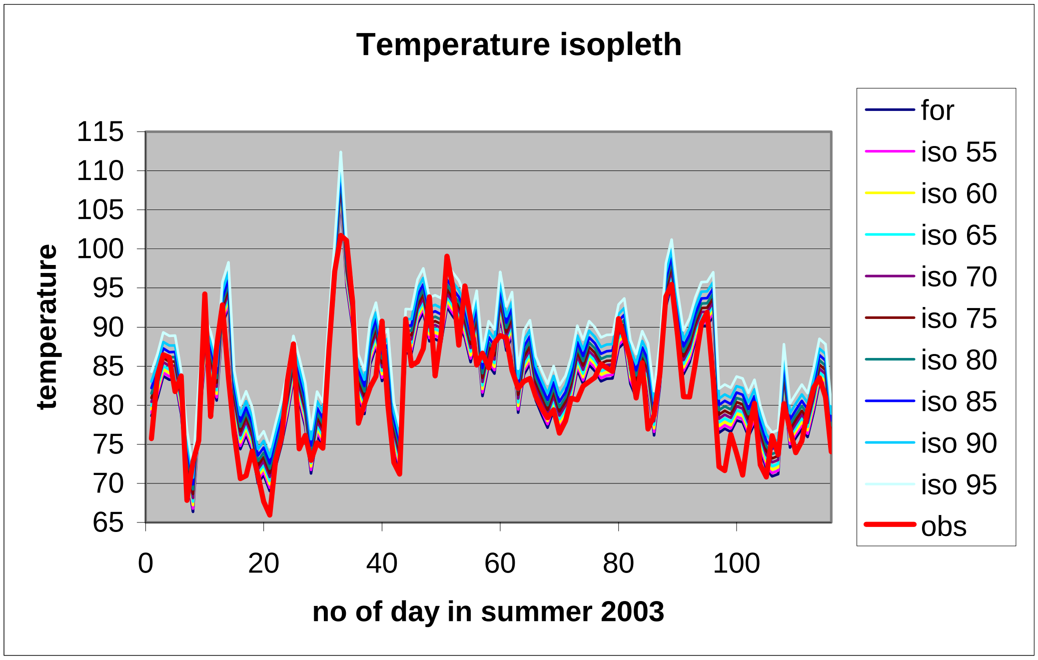


Figure 2-12. The Distribution of Forecast Temperature Isopleths Using PF1 for the Bay Area Over the Summer of 2003.
Only days with full forecast information are shown). The observed Cal ISO region temperature is shown in red. Note that by playing a higher-ranking isopleth allows a better upper bound on the temperature.

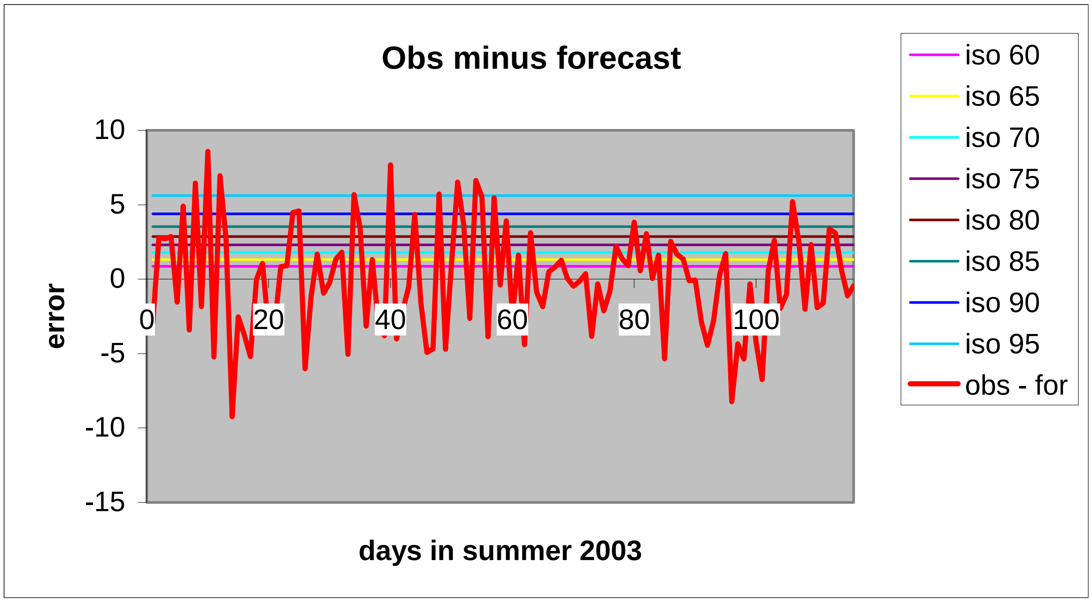


Figure 2-13. As Figure 12, But Showing Variations in the Observed Temperature Within the Isopleths

Full California Economic Impact Forecast Results

In this section we translate the temperature forecasts into relative economic impact in millions of dollars. (a 50% saving would reflect a real reduction of weather-related costs by half.). The results below have been computed by translating the temperature forecast error in each region into a megawatt error (with the kind assistance of Dennis Gaushell), and then the four region megawatt error into an economic cost for each day using the cost curve of **figure 2-5**. The target weather-related demand is computed using the observed temperatures.

One Day Ahead Forecasts

The table below contains the economic costs of using the 50%, 55%, 60%,…isopleth for each of the probability forecast packages described in the sections above for the 2003 summer season; only days with full forecast availability were used.

(table rescinded)

The best performance is for the 55% isopleth of PF1, note however that the same isopleth of AVN does quite well. These two forecasts score $X and $X respectively, significantly below the Cal ISO AVN value of $X for the season. Cal ISO averages only three forecasts, the AVN, ETA and MRF. Thus, their average of the three is X. Also, this method allows an improvement in using the AVN forecast results as is, which is immediately useful to CalISO as this is the model they operationally use.

The likely uncertainty of these results has been quantified by bootstrap resample of the days (keeping the weekday as weekdays in this case). The table below reproduces the results of the previous table for the 50, 55, 60, 65, and 70% isopleths, while including the standard deviation estimated with via the bootstrap. The results between models are significant, while the choice between the 60, 55 and 50% isopleths is less clear. To investigate the origin of this we consider scatter graphs of the forecasts in detail.

(table rescinded)

Figures 2-14 and **2-15** contrast the cost of the current Cal ISO forecast with that of the 55% isopleth of our AVN probability model, points above the diagonal indicate days on which the AVN 55 line outperforms the Cal ISO AVN. (Note: units normalized in rescinded version.)

Figure 2-14 shows all the data including two days with major savings, while **figure 2-15** contains a zoom near the origin which shows that the savings from the AVN 55 line are significant on 'normal' days as well as. Note that it is not uniformly better, there are days where the other method has a lower cost, but on average the 55% isopleth of AVN is clearly better.

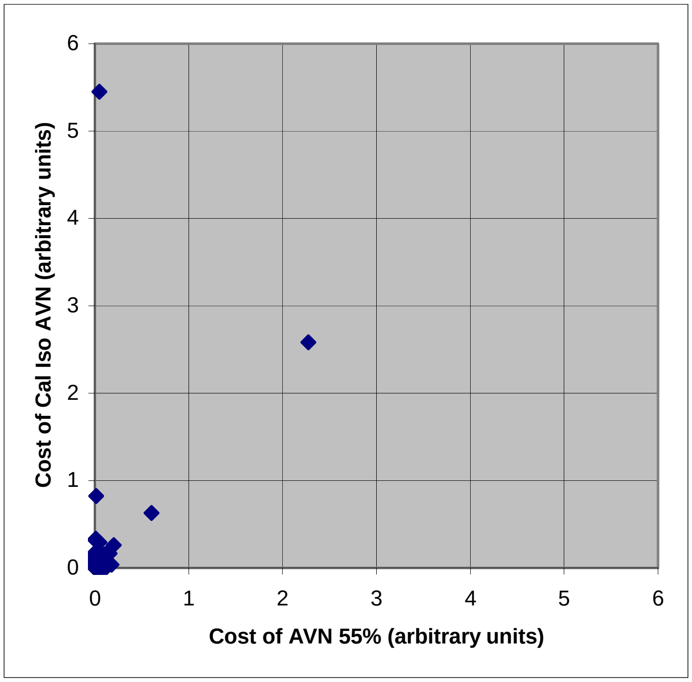


Figure 2-14. Scatter Diagram of Cal ISO AVN Against the 55% Isopleth of Our AVN; Each Point Represents One Day.
The full range of the data are shown.

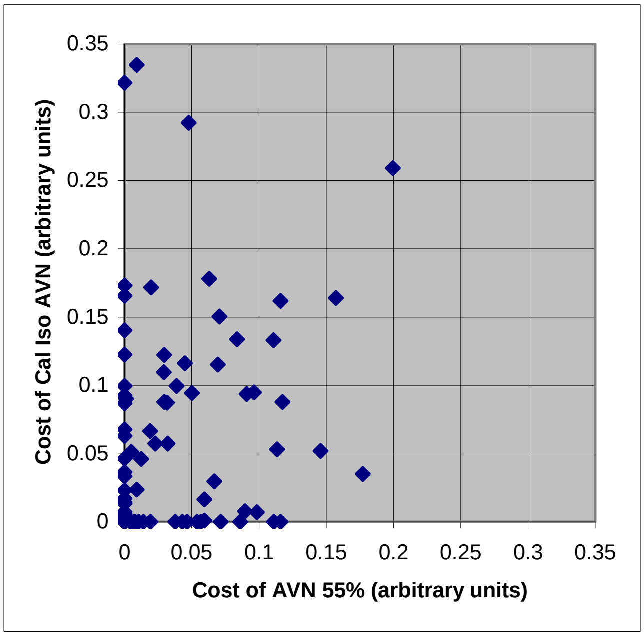


Figure 2-15. Scatter Diagram of Cal ISO AVN Against the 55% Isopleth of Our AVN;
Each Point Represents One Day.
A zoom near the origin is shown; Note that there are many days where one or the other of the forecasts scores zero cost.

Finally we contrast the 50% isopleth of our AVN model with the 55% isopleth of the same model. Note on the large scale graph (**figure 2-16**) that again there are costly points well above the diagonal, indicating that the 55% isopleths does better on these days. The cost for this is over supply on days when the observed temperature is near the forecast mean. This is more clearly seen in **figure 2-17**, which shows a zoom.

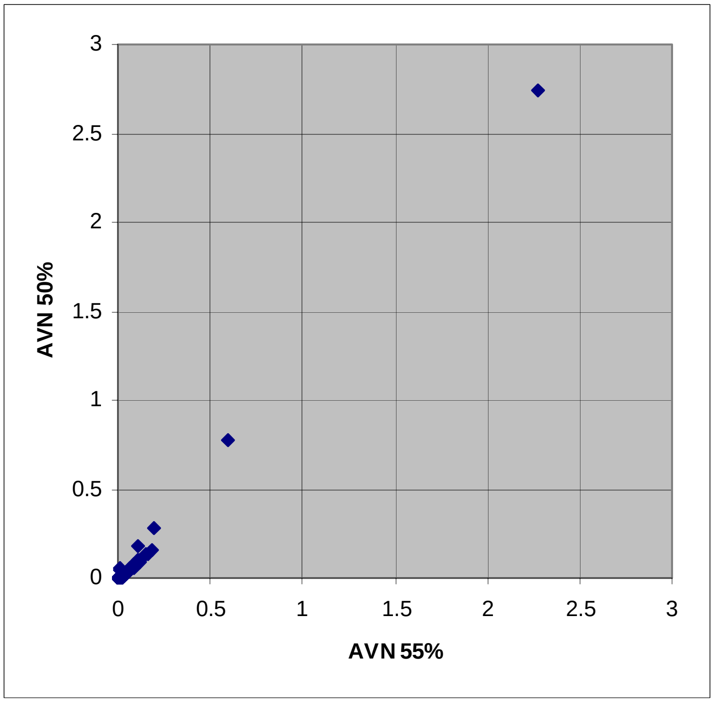


Figure 2-16. Scatter Diagram of 50% Isopleth Against the 55% Isopleth of Our AVN

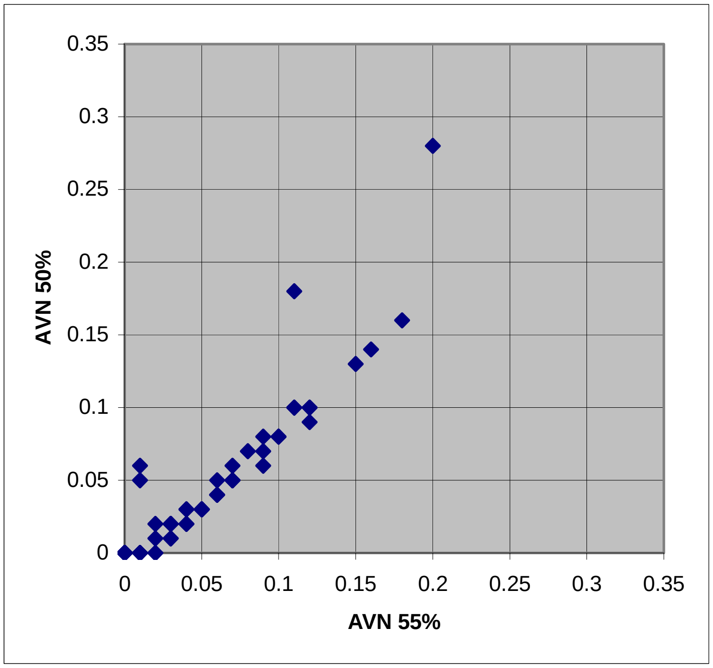


Figure 2-17. Scatter Diagram of the 50% Isopleth Against the 55% Isopleth of Our AVN; Each Point Represents One Day.
This graph is a zoom near the origin.

Note in **figure 2-17** that many days lie below the diagonal. Thus on the major under forecast days the 55% isopleth pulls significantly ahead. Four examples are shown in **figure 2-17**, a few even larger deviations are shown in **figure 2-16**.

Two Day Ahead Forecasts

Moving to two day ahead forecasts requires a new statement of the components which go into the PF models,

Package No	Package Name	No	Model	Number of predictors	Model run time 1	Delay 1	UTC time 1	Model run time 2	Delay 2	UTC time 2
1	AVN	1	avn	2	42	0	18	45	0	21
		2	avn_mos	2	42	0	18	45	0	21
		3	avnext	0						
2	ENS	4	ens	1	48	0	0			
		5	ens_max	1	48	0	0			
		6	ens_mean	1	48	0	0			
		7	ens_min	1	48	0	0			
		8	ens_sdn	1	48	0	0			
		9	ens_sdp	1	48	0	0			
3	ETA	10	eta	2	42	0	18	45	0	21
		11	eta_mos	2	42	0	18	45	0	21
4	MEM	12	mem	1	48	0	0			
		13	mem_max	1	48	0	0			
		14	mem_mean	1	48	0	0			
		15	mem_min	1	48	0	0			
		16	mem_sdn	1	48	0	0			
		17	mem_sdp	1	48	0	0			
5	MRF	42	mrf	1	48	0	0			
		19	mrf_mos	1	48	0	0			
6	NMGMOS	20	ngm_mos	2	42	0	18	45	0	21
7	ENSMEAN	21	ens_mean	1	48	0	0			
		22	ens_sdp	1	48	0	0			

The corresponding temperature forecasts for each region and each Package are shown in the following 4 tables, again PF1 and our AVN package show the best performance, although the other PF packages, including PF% sometimes score well.

PG&E Bay Area Summer 2003 48 Hours							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.00	4.63	-0.71	0.09	4.43	0.23
2	ENS	0.00	7.25	-0.06	0.07	1.72	0.06
3	ETA new	0.00	5.02	-0.59	0.09	3.71	0.17
4	MEM	0.00	6.94	-0.12	0.05	1.88	0.07
5	MRF	0.00	6.05	-0.32	0.08	2.48	0.14
6	NGMMOS	0.00	4.68	-0.69	0.10	4.21	0.21
7	ENS new	0.00	6.42	-0.24	0.06	2.21	0.09
8	PF1	0.01	4.29	-0.82	0.11	4.96	0.23
9	PF2	-	-	-0.56	0.04	2.65	0.09
10	PF3	-	-	-0.60	0.04	2.82	0.10
11	PF5	-	-	1.46	0.62	14.15	2.21
12	Cal ISO	-	-	-	-	-	-

PG&E Non Bay Area Summer 2003 48 Hours							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.01	2.59	-1.38	0.06	13.62	0.51
2	ENS	0.01	4.28	-0.65	0.08	5.02	0.24
3	ETA new	0.01	2.95	-1.19	0.08	11.01	0.46
4	MEM	0.00	4.56	-0.56	0.10	4.44	0.23
5	MRF	0.01	4.03	-0.74	0.09	5.48	0.31

6	NGMMOS	0.01	3.14	-1.10	0.09	9.46	0.41
7	ENS new	0.01	3.99	-0.76	0.07	5.75	0.21
8	PF1	0.01	2.61	-1.37	0.07	13.96	0.54
9	PF2	-	-	-1.09	0.04	6.96	0.25
10	PF3	-	-	-1.15	0.04	7.58	0.26
11	PF5			-1.27	0.10	19.99	1.41
12	Cal ISO			-	-	-	-

San Diego Edison Summer 2003 48 Hours							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.01	3.33	-0.97	0.09	8.76	0.36
2	ENS	0.00	4.92	-0.41	0.10	3.94	0.23
3	ETA new	0.01	3.66	-0.84	0.10	7.23	0.37
4	MEM	0.00	4.73	-0.46	0.08	4.34	0.19
5	MRF	0.01	4.20	-0.64	0.07	5.38	0.20
6	NGMMOS	0.01	3.35	-0.96	0.10	8.93	0.48
7	ENS new	0.01	4.23	-0.63	0.10	5.33	0.28
8	PF1	0.01	3.18	-1.04	0.09	9.76	0.39
9	PF2	-	-	-0.86	0.07	5.76	0.26
10	PF3	-	-	-0.89	0.07	6.02	0.25
11	PF5	-	-	-0.88	0.19	16.49	1.79
12	Caliso	-	-	-	-	-	-

San Diego Gas and Electric Summer 2003 48 Hours							
Number	Package	Bias	Standard Deviation Error	Mean rel. Ignorance	Bootstrap Deviation Ignorance	Brier Score (*1000)	Bootstrap Deviation Brier Score
1	AVN	0.01	4.24	-0.75	0.10	5.17	0.35
2	ENS	0.00	5.18	-0.46	0.08	3.29	0.19
3	ETA new	0.00	4.49	-0.67	0.09	4.53	0.24
4	MEM	0.00	4.89	-0.55	0.07	3.66	0.20
5	MRF	0.00	4.89	-0.55	0.08	3.76	0.18
6	NGMMOS	0.00	4.38	-0.71	0.07	4.86	0.23
7	ENS new	0.00	4.98	-0.52	0.07	3.52	0.17
8	PF1	0.01	3.95	-0.86	0.07	6.06	0.23
9	PF2	-	-	-0.70	0.06	3.86	0.17
10	PF3	-	-	-0.71	0.05	3.90	0.16
11	PF5	-	-	-0.34	0.30	11.17	0.90
12	Caliso	-	-	-	-	-	-

The minimum cost isopleth for day 2 is at 65%, this represents moving even further out on the tail of the distribution in the attempt to avoid the cost of under-forecasting. These results are based on the 137 days in the 2003 season for which the forecasts could be build and analyzed. Bootstrap standard deviations are determined from 1000 resamplings.

(costs rescinded)										

The next two graphs, **figures 2-18** and **2-19**, again plot the weather related cost on a given day ‘best’ isopleth (in this case 65%) against the cost on the same day of using the 50% isopleth. The outlier is the 28 of May, which has been removed on the zoom. The same pattern is evident that appeared on the one day case. The strategy accepts regular, relatively small, loses but then mitigates the occasional large loss. Note from the 65STD column in the table that the uncertainty in skill is fairly large: several of the PF packages and the AVN package are well within the bootstrap resampling standard deviation, the statistics being dominated by a few days with very high potential losses.

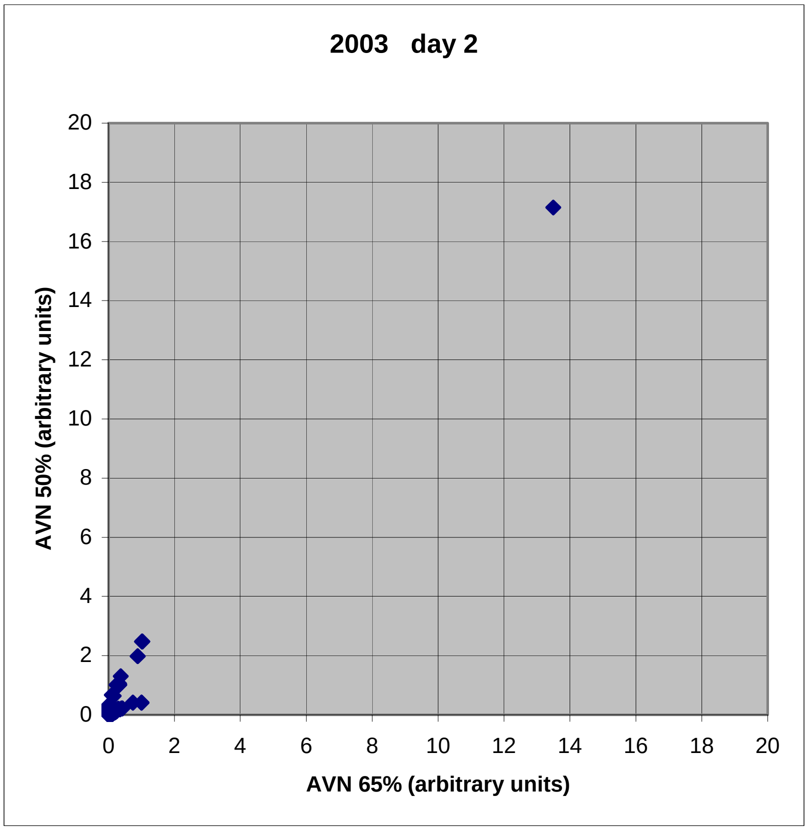


Figure 2-18. Scatter Diagram of AVN 50% Against the AVN 65%

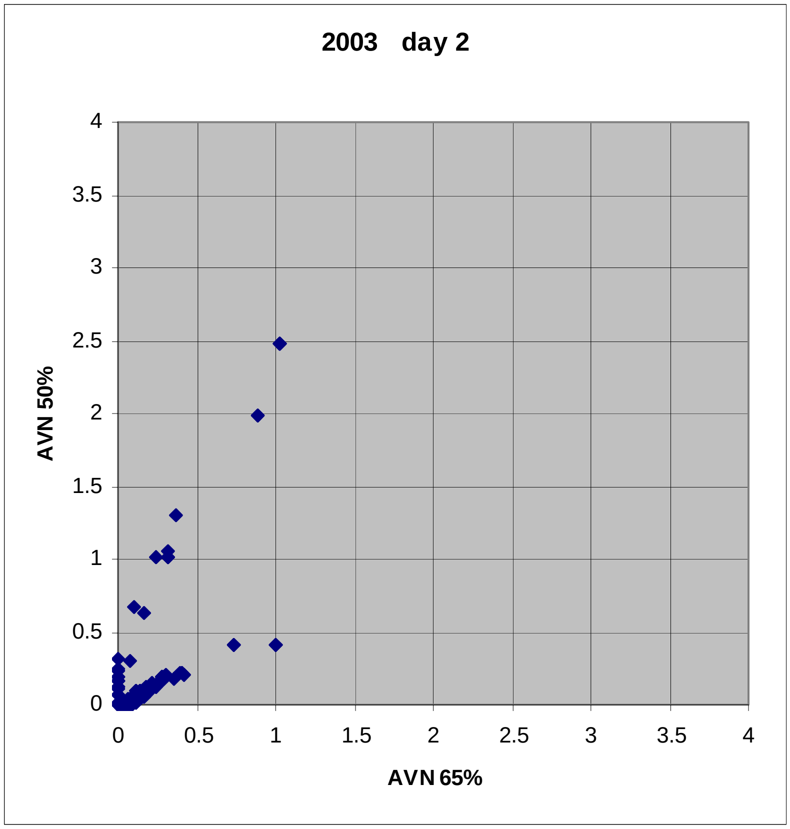


Figure 2-19. Scatter Diagram (Zoom) of AVN 50% Against the AVN 65%

Forecasts and Errors of 2002 with AVN – SVD Fitted to 2002 and 2003

This section shows that the predictions based on a different year are slightly less accurate than predictions based on the same year. 2002 and 2003 are contrasted, first forecasting 2002 with models built on the 2002 data, and one build on the 2003 data. Based on these results, it appears that a combination of the previous summer data and the current summer (starting, for example, in April) can be used. The current year April data should be monitored carefully to see if there have been any major changes in any of the models. This monitoring should continue throughout the summer, the model being updated daily to include the new observations and corresponding forecast (that is, the parameters are refit each day).

Forecasting 2002	RMSE	BIAS
Model from 2002	1.59	0.00
Model from 2003	3.42	0.82

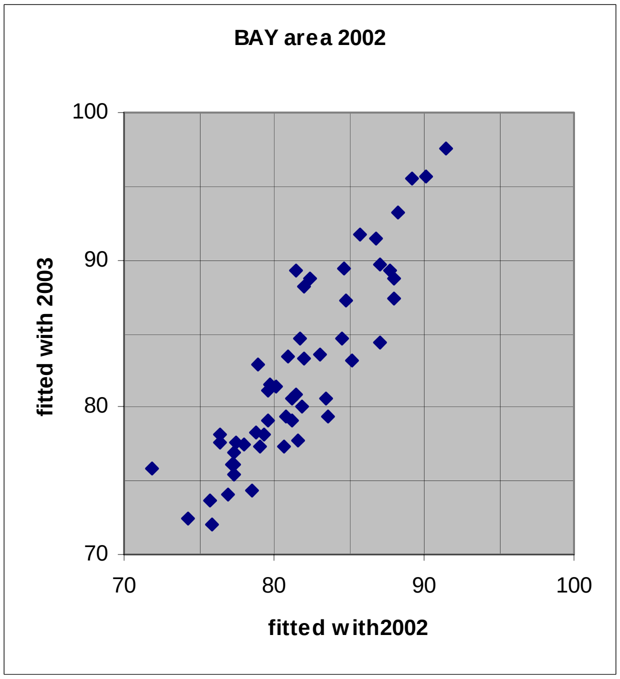


Figure 2-20. 50% Isopleth Prediction for Bay Area in 2002

Figure 2-20 shows the 50% isopleth predictions made for each day in 2002. The x-axis is the prediction using the 2002 model and the y-axis that using the 2003 model. For smaller temperatures, the predictions scatter about the diagonal, which indicates the model(s) generalise from one summer to another. For the higher predicted temperatures, however, the 2003 model is consistently higher than the 2002 model. Given there are only 54 days with complete data, it is difficult to break these results down into further sub-categories. **Figure 2-21** shows a zoom of **figure 2-20**.

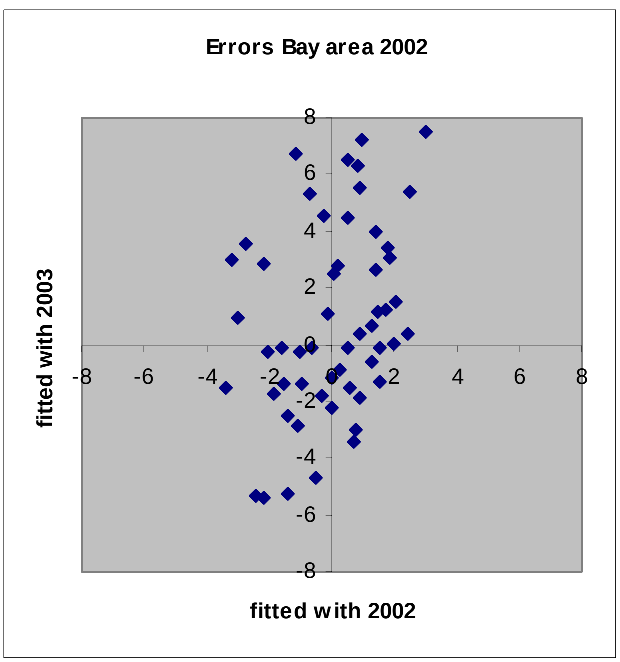


Figure 2-21. 50% Isopleth Prediction (Zoom) for Bay Area in 2002

Forecasts of 2003 with AVN – SVD fitted to 2002 and 2003

The table and graphs (**figures 2-22** and **2-23**) below repeat this analysis, but this time the forecasting of the 2003 data from models built using the 2003 data (drop-one-out) and a model based on all the 2002 data.

Forecasting 2002	RMSE	BIAS
Model from 2002	4.17	-0.63
Model from 2003	3.35	-0.01

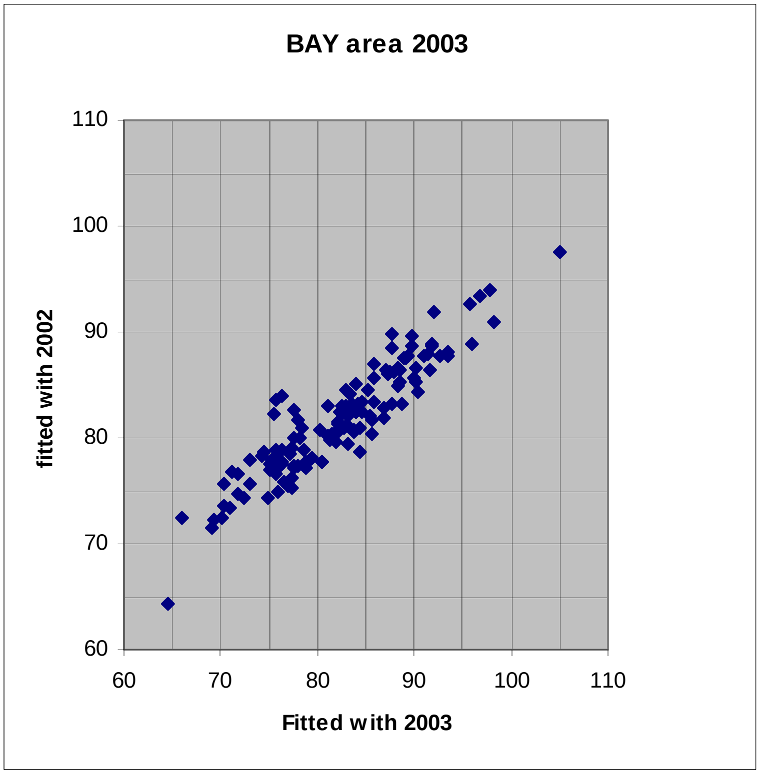


Figure 2-22. 50% Isopleth for Bay Area 2003

Note the model-2003 systematically forecasts more extreme values than model-2002. This is consistent with the negative bias of model-2002.

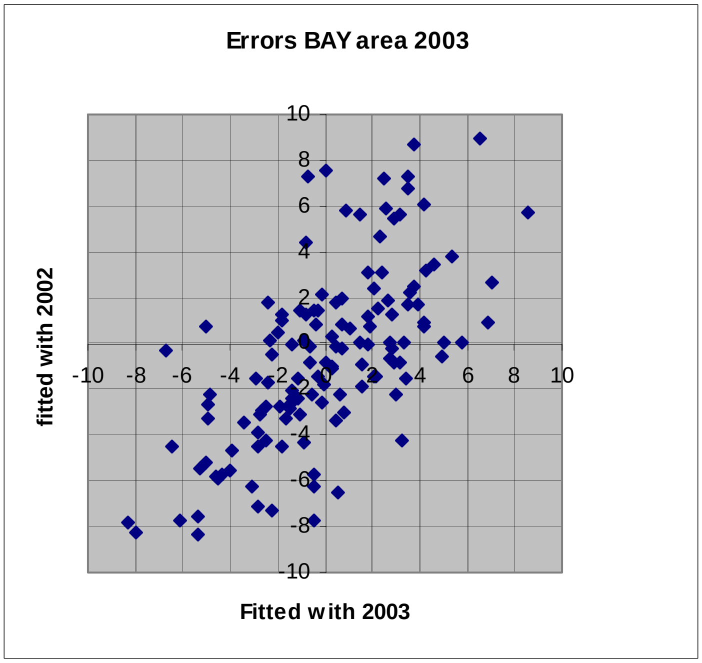


Figure 2-23. 50% Isopleths (Zoom) for Bay Area 2003

Observed demand versus 1,2,3 day forecast demand by month.

The graphs below show, on the same grid, one, two and three day forecasts. The aim is to allow an operator to see what would have happened if they had made a demand forecast a few days earlier, as for example over the weekend. The weekends are spotted as the low demand days; not that on the Bay area Mondays, the 3 day forecasts is rather similar to the one day in the May/June period shown, whereas the three day forecasts appear worse, especially for the AVN, in July. Note that for the one day lead time forecasts the 55% isopleth is used, while at days two and three the 65% isopleth is used (in each case, we use the optimal isopleth for that model for that lead time optimised over the entire summer season of the same year.)

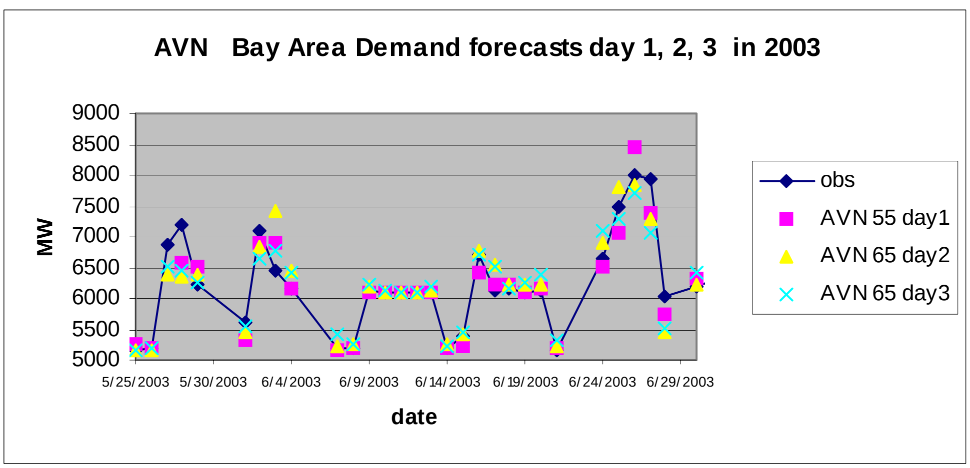


Figure 2-24. May June 2003 AVN

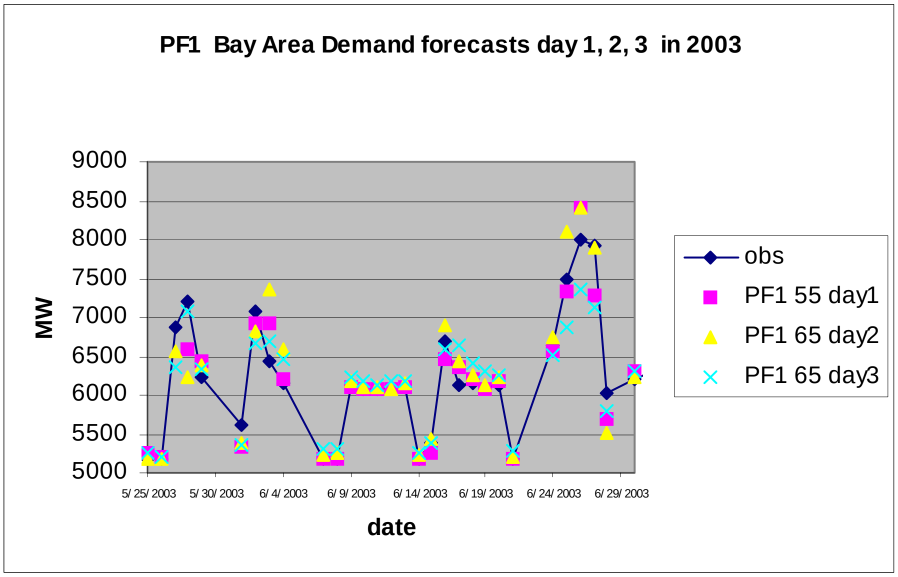


Figure 2-25. May-June 2003 PF1

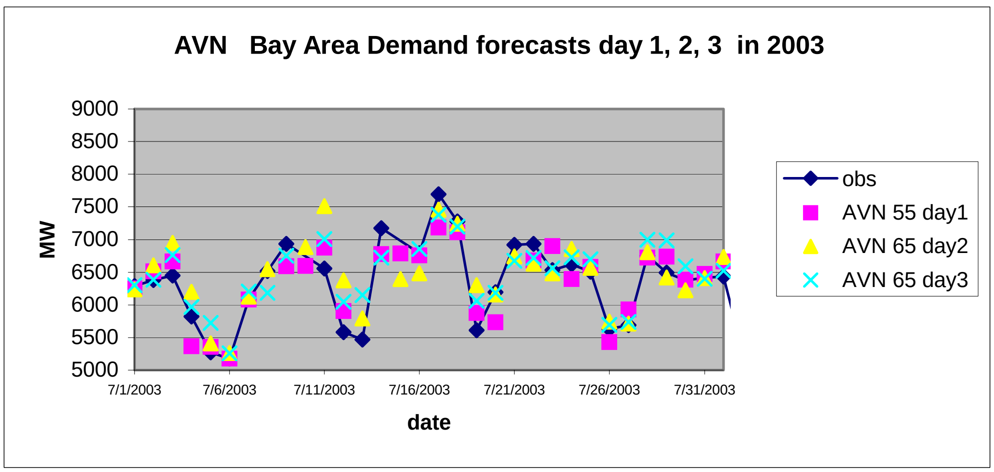


Figure 2-26. July 2003 AVN Bay Area

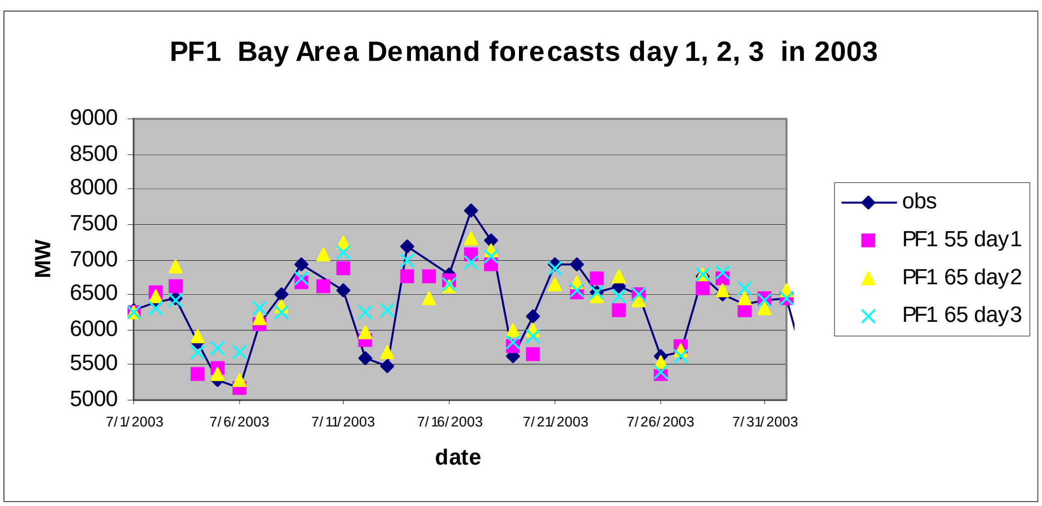


Figure 2-27. July 2003 PF1 Bay Area

Function i) from Temperature to Demand (MW)

The following is a piece of code based upon the formulas Cal ISO uses. Note that demand is a quadratic function of temperature with different parameters in the four regions, as well as distinguishing workdays and holidays. This code translates from temperature to demand.

```
!*****************************************************************
! Function which takes the temperature to the demand as a function
! of workday (called weekday) and the region
! input: temp, weekday, region
! output: demand
!*****************************************************************
    function demand(region, weekday, temp)
    implicit none
    real*8 temp, demand
    character*4 region
    logical weekday
    if(region .eq. "BAY")then
       if(weekday .eq. .true.)then
           demand=  1.8537*temp*temp - 257.97*temp+15056
       else
           demand = 2.1336*temp*temp - 313*temp+16653
       endif

    elseif(region .eq. "NBAY")then
       if(weekday .eq. .true.)then
           demand= 3.3722*temp*temp  - 461.04*temp+23899
       else
           demand = 3.5316*temp*temp - 493.02*temp+24512
       endif

    elseif(region .eq. "SCE")then
       if(weekday .eq. .true. )then
           demand= 6.8592*temp*temp -969.02*temp+46432
       else
           demand = 5.5899*temp*temp -812.57*temp+40379
       endif

    elseif(region .eq. "SDGE")then
       if(weekday .eq. .true. )then
           demand = 1.0445*temp*temp-141.85*temp+7477
       else
           demand= 0.8491*temp*temp-121.17*temp+6757
       endif
    endif
    return
    end
```

Ignorance as a Function of Time

Figures 2-28 and **2-29** show the ignorance of the worst and the best regions, specifically BAY and NBAY respectively. It is interesting to note that PF5 performs very poorly in the Bay area, having positive ignorance (that is, worse than climatology) from day two, while in the non-Bay region PF5 performs similarly to PF3.

Overall, note that PF1 does well (if not best) in terms of ignorance score in all cases, while AVN is similarly good.

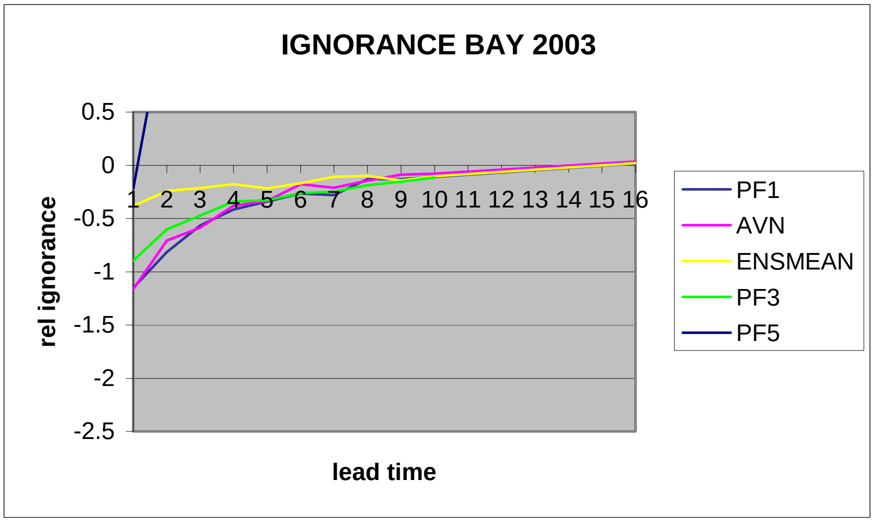


Figure 2-28. Ignorance of Bay 2003

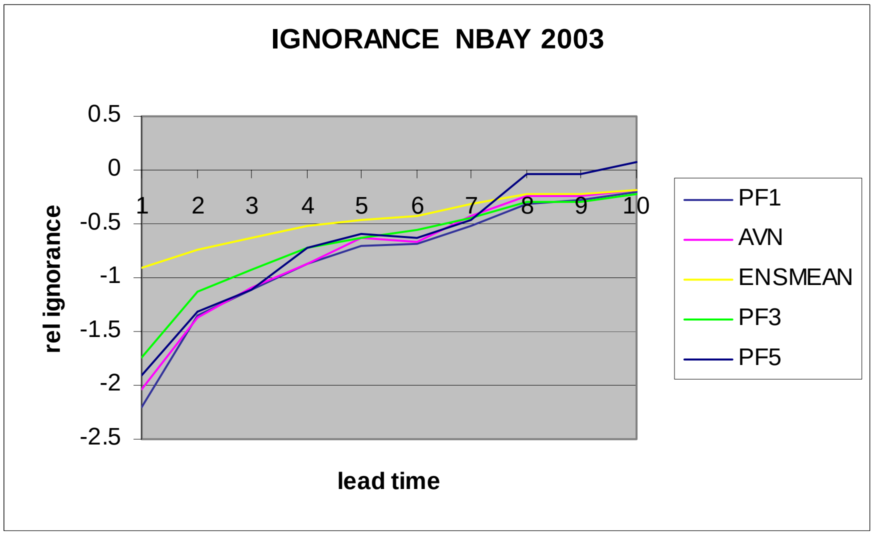


Figure 2-29. Ignorance of Non Bay 2003

RMSE as a Function of Time

Figures 2-30 and **2-31** show the RMSE of the worst and the best areas (BAY and NBAY). Models PF3 and PF5 do not produce an RMSE error (they are explicitly distribution forecast models). It is interesting to note that the NCEP IC-ensemble mean is NOT significantly better that the other forecasts. Note that PF1 scores about the same as the AVN Package in the Bay Region, but systematically better in the non-Bay region at longer lead times.

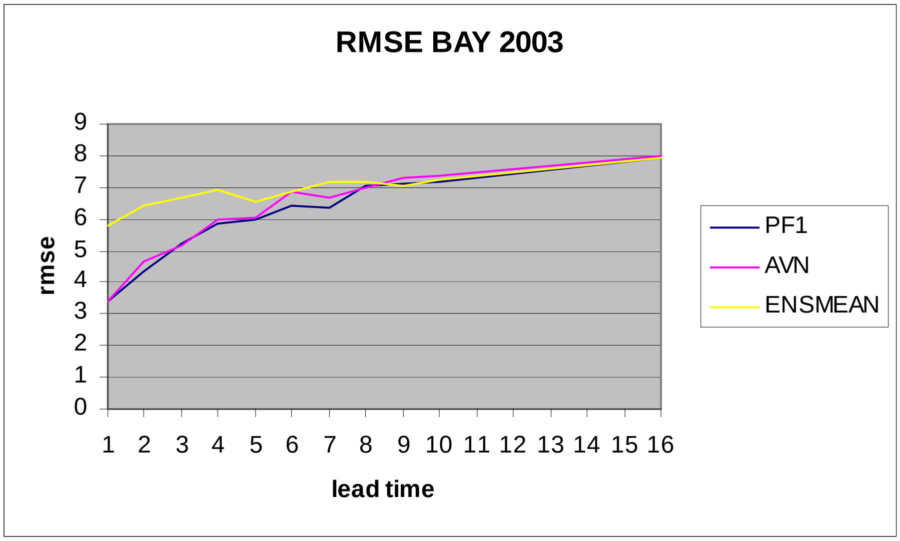


Figure 2-30. RMSE of Bay

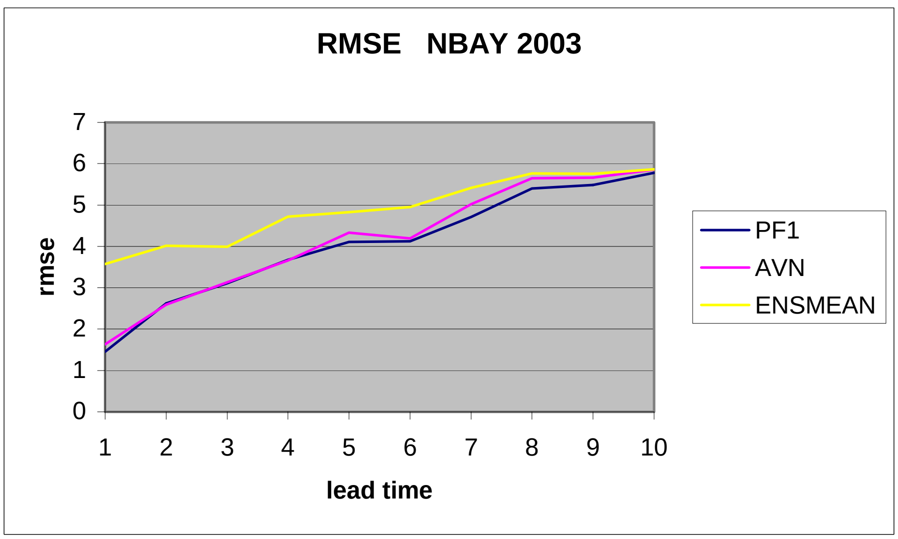


Figure 2-31. RMSE of Non Bay

Optimal Quantile as a Function of Time, for Lead Times One to Three

The table below shows the optimal quantile (OQ), specifically the isopleth that minimizes the cost over the summer of 2003 (thus it is not out-of-sample). The larger the value, the farther out on the tail the optimal isopleth is located; under each model this quantity usually increases with lead time, that is, the further in the future the higher the optimal isopleth is from the median.

Note, this is derived from the cost which is averaged over all four areas

Lead	OQ AVN	OQ PF1	OQ PF3	OQ PF5
1	55	55	60	70
2	65	65	70	70
3	65	65	70	85

Offset (Safety Factor) in Temperature as a Function of Time

The optimal quantile corresponds to a temperature greater than that of the median (that is, the 50% isopleth). The difference between the temperature at the optimal isopleth and that at the 50% isopleth is effectively a safety factor that serves as temperature excess on most days. For the AVN package and PF1 this is a constant value; for PF3 and PF5 this value changes from day to day (and the number in the table below is an average over the 2003 summer dates.)

Note, this is just for the BAY area.

Lead	AVN	PF1	PF3 (average value)	PF5 (average value)
1	0.43	0.43	1.00	3.05
2	1.79	1.67	2.67	3.36
3	2.00	2.03	2.85	6.95

Three Tables showing cost by model by quantile for day 1, 2 and 3.

(table rescinded)

(table rescinded)

(table rescinded)

The minimum cost (in unknown units of dollars) per model is shown in magenta for each model. The absolute minimum over all models is shown in yellow. Note that AVN-Package wins for day 2 and 3, for day 1 PF1. Note, the NGM model is not available for the 3 day forecast.

Economic Benefits Relative to Costs of Procuring Operational Ensemble Predictions

The expected benefits of the forecasts of moving to ensemble methods are substantial.

They carry with them two increased costs: First the full forecast suite must be obtained, not merely the AVN forecast. Secondly, significantly more information processing must be done on these inputs. In addition to the direct costs, there is the additional operational requirement for robust response in case one or more of the forecast models is not available on a given day.

Three forecast information providers gave estimated costs for the full model suite required by PF1. Expected prices varied a great deal, especially regarding volume discounts, but the range of values for 20 stations, full forecast suite two times per day, over a period of six months ranged from about $5,000 to $50,000. Even the high end of this range is significantly less than the expected savings from reduced weather related demand forecast errors.

Conclusion

There is clearly significant benefit in using existing ensemble of weather forecast models towards significantly reducing Cal ISO exposure to weather related demand forecast errors; relative the current method (here after called the standard method) of using the AVN T_max forecasts the weather-related costs of the 2003 test period can be cut in half.

This improvement comes from two sources. Firstly weather information is not taken at face value, but a number of forecasts from each model are combined, via singular value decomposition, to determine a better point forecast based on that model. Hence our AVN_Package forecast used more input information than merely the AVN T_max, and hence it too significantly outperforms the standard method, which is also based on AVN output. Secondly, our approach constructs a forecast distribution, not a single number. This allows us to incorporate Cal ISO's highly asymmetric cost function in a systematic way. Each distribution consists of levels, called isopleths. By determining empirically which isopleth yields the best balance between overproduction and underproduction, knowledge of uncertainty in the ensemble forecasts is used to systematically set the "safety margin". This approach allows one to quantify weather related demand forecast confidence in a more quantitative and automatic manner at a relatively low cost.

A multi-model ensemble using the 55% isopleth of the forecast distribution provided the most useful forecasts. This model takes as input several variables at each station from each of the numerical weather prediction models available from quantumweather. The cost of the raw inputs varies with supplier, but is several orders of magnitude less than the savings based on the 2003 summer data.

These results are significant under bootstrap testing, and cross validation on a limited amount of data from 2002 (that is, using the model constructed using 2003 data to forecast 2002 demand) indicated that one should expect skill in true out-of-sample demand forecasting. All initial model selections were done under drop-one-out cross-validation (rebuilding the model for the entire summer to forecast each day so that data on this date is not included when forecasting that day) and the bootstrap significance levels are quoted in the tables.

This Page Intentionally Left Blank

Appendix 1: Target Weather Stations and Weights

The table below summarizes current data on the stations originally defined by Cal ISO. The columns indicate whether or not forecast products for that station are available from QuantumWeather, which alternative stations have been identified if not forecast products are available; a description of the station, its location and WMO identifier; and whether or not observations for that station were included on the "CAL ISO data 2/18/2004" CD.

WEATHER STATIONS AND WEIGHTS FOR CAL ISO LOAD FORECASTING

Code	QW Data?	Alternative Station	Description	Weight	Lat	Lon West	(0-360)	WMO ID	Feb D
PG&E Bay Area									
KSFO	yes		San Francisco Airpo	0.05	37.62	122.38	237.62	724940	yes
KOAK	yes		Oakland Airport	0.05	37.73	122.2	237.8	724930	yes
KCCR	yes		Concord	0.27	37.98	122.05	237.95	724936	yes
KLVK	yes		Livermore	0.27	37.7	121.82	238.18	724927	NO
KSJC	yes		San Jose Airport	0.36	37.37	121.92	238.08	724945	NO
PG&E No Bay Area									
KSTS	yes		Santa Rosa	0.05	38.52	122.82	237.18	724957	yes
KPRB	yes		Paso Robles	0.05	35.67	120.63	239.37	723965	yes
KRDD	yes		Redding	0.09	40.5	122.3	237.7	725920	yes
KMYV	yes		Marysville	0.27	39.1	121.57	238.43	724838	yes
KMCE	yes	NoMOS	Merced	0.27	37.28	120.51	239.49	724815	yes
KBFL	yes		Bakersfield	0.27	35.42	119.05	240.95	723840	yes
SCE									
KCVC	NO	KLAX KFUL	Civic Center	0.165				?	
KFUL	yes		Fullerton	0.165	33.87	117.97	242.03	722976	NO
KONT	yes		Ontario	0.23	34.05	117.62	242.38	722865	NO
KPSP	yes		Palm Springs	0.22	33.83	116.5	243.5	722868	NO
KWJF	yes		Lancaster	0.22	34.73	118.22	241.78	723816	NO
SDGE									
KOKB	NO	KL32 KL34 KNFG	Oceanside	0.1	33.22	117.35	242.65	722934	NO
KSAN	yes		Lindbergh Field	0.15	32.73	117.17	242.83	722900	NO
KRNM	NO	KNJK KSEE	Ramona	0.375	33.04	116.92	243.08	745056	NO
KSEE	yes	NoMOS	Santee/Gillespie	0.375	32.82	116.97	243.03	722907	NO

Statements of input variables for AVN and PF1

AVN

The AVN package uses the 18 and 21 hour lead time forecasts verifying on the next day at 18:00 and 21:00 UTC (verifying California local time 10:00 and 13:00) from the two 'models' called AVN and AVN_MOS.

Package No	No	Model	Number of Predictors	Model run time 1	Delay 1	UTC time 1	Model run time 2	Delay 2	UTC time 2
1	1	avn	2	18	0	18	21	0	21
	2	avn_mos	2	18	0	18	21	0	21

PF1

PF1 uses as input the forecasts from packages 1-6. In order to produce statistical forecasts for the 6 packages the following numerical forecast products are needed:

The AVN package: 18 and 21 hour lead time forecasts verifying on the next day at 18:00 and 21:00 UTC (verifying California local time 10:00 and 13:00) from the two 'models' called AVN and AVN_MOS.

The ENS package: the 24 hour lead time forecast verifying 0:00 UTC time next day (= 16:00 local time) of the control, mean, max, min, mean plus and minus a standard deviation, or the whole ensemble.

The ETA package: 18 and 21 hour lead time forecasts verifying on the next day at 18:00 and 21:00 UTC (verifying California local time 10:00 and 13:00) from the two 'models' called ETA and ETA_MOS.

The MEM package: the 24 hour lead time forecast verifying 0:00 UTC time next day (= 16:00 local time) of the control, mean, max, min, mean plus and minus a standard deviation, or the whole ensemble.

The MRF package: the 24 hour lead time forecast verifying 0:00 UTC time next day (= 16:00 local time) of the two models called MRF and MRF_MOS.

The NGM package: 18 and 21 hour lead time forecasts verifying on the next day at 18:00 and 21:00 UTC (= 10:00 and 13:00 local time) from the 'model' called NGM_MOS.

Package No	No	Model	Number of Predictors	Model run time 1	Delay 1	UTC time 1	Model run time 2	Delay 2	UTC time 2
1	1	avn	2	18	0	18	21	0	21
	2	avn_mos	2	18	0	18	21	0	21
	3	avnext	0						
2	4	ens	1	24	0	0			
	5	ens_max	1	24	0	0			
	6	ens_mean	1	24	0	0			
	7	ens_min	1	24	0	0			
	8	ens_sdn	1	24	0	0			
	9	ens_sdp	1	24	0	0			
3	10	eta	2	18	0	18	21	0	21
	11	eta_mos	2	18	0	18	21	0	21
4	12	mem	1	24	0	0			
	13	mem_max	1	24	0	0			
	14	mem_mean	1	24	0	0			
	15	mem_min	1	24	0	0			
	16	mem_sdn	1	24	0	0			
	17	mem_sdp	1	24	0	0			
5	18	mrf	1	24	0	0			
	19	mrf_mos	1	24	0	0			
6	20	ngm_mos	2	18	0	18	21	0	21

List of stations used

PG&E Bay Area	PG&E No Bay Area	SCE	SDGE
KSFO	KSTS	KLAX	KNFG
KOAK	KPRB	KFUL	KSAN
KLVK	KRDD	KONT	KNJK
KSJC	KMYV	KPSP	KSEE
KCCR	KMCE	KWJF	
	KBFL		

Appendix 2: Model Sources and Characteristics

	Source	Weather Model Name	Old Model Name	Intra-day runtimes	Forecast Variable	Forecast time step	Forecast Duration	Web Description
1	NWS	ETA		0z, 6z, 12z, 18z,	2m Temp	3hrs	84 hrs	http://www.emc.ncep.noaa.gov/modelinfo/index.html
2	NWS	ETA MOS		0z,12z	2m Temp	3 hrs beyond hr 6	72 hrs	http://www.nws.noaa.gov/mdl/synop/metcard.htm
3	NWS	NGM MOS		0z,12z	2m Temp	3 hrs beyond hr 6	60 hrs	http://www.nws.noaa.gov/mdl/synop/fwcexpln.htm
4	NWS	GFS MOS	AVN MOS	0z, 6z, 12z, 18z,	2m Temp	3 hrs beyond hr 6	72 hrs	http://www.nws.noaa.gov/mdl/synop/mavcard.htm
5	NWS	GFS	AVN	0z, 6z, 12z, 18z,	2m Temp	3hrs up to 180, 12 hrs beyond 180	384 hrs	http://www.emc.ncep.noaa.gov/modelinfo/index.html
6	NWS	GFS Ensemble-Control (or simply ENS)	MRF ENS	0z,12z	2m Temp	12 hrs	384 hrs	http://www.emc.ncep.noaa.gov/modelinfo/index.html
7	NWS	GFS Ensemble-Mean		0z,12z	2m Temp	12 hrs	384 hrs	
8	NWS	GFS Ensemble- +1 std		0z,12z	2m Temp	12 hrs	384 hrs	
9	NWS	GFS Ensemble- -1std		0z,12z	2m Temp	12 hrs	384 hrs	
10	NWS	GFS Ensemble-Max		0z,12z	2m Temp	12 hrs	384 hrs	
11	NWS	GFS Ensemble-Min		0z,12z	2m Temp	12 hrs	384 hrs	
12	NWS	GFS Ensemble MOS Control (or simply ENS MOS)	MRF ENS MOS	0z	2m Temp	12 hrs beyond hr 24	192 hrs	http://www.nws.noaa.gov/mdl/synop/enstxt.htm
13	NWS	GFS Ensemble MOS Mean		0z	2m Temp	12 hrs beyond hr 24	192 hrs	
14	NWS	GFS Ensemble MOS +1 std		0z	2m Temp	12 hrs beyond hr 24	192 hrs	
15	NWS	GFS Ensemble MOS - 1std		0z	2m Temp	12 hrs beyond hr 24	192 hrs	
16	NWS	GFS Ensemble MOS Max		0z	2m Temp	12 hrs beyond hr 24	192 hrs	
17	NWS	GFS Ensemble MOS Min		0z	2m Temp	12 hrs beyond hr 24	192 hrs	
18	NWS	MRF		0z	2m Temp	12 hrs	384 hrs	http://www.emc.ncep.noaa.gov/modelinfo/index.html
19	NWS	GFSX MOS	MRF MOS	0z	2m Temp	12 hrs beyond hr 24	192 hrs	http://www.nws.noaa.gov/mdl/synop/mexcard.htm

This Page Intentionally Left Blank