Davison Van Cleve PC
Attorneys at Law
TEL (503) 241-7242 • FAX (503) 241-8160 • mail@dvclaw.com
Suite 400
333 SW Taylor
Portland, OR 97204

August 13, 2010

Via Electronic and US Mail

Public Utility Commission
Attn: Filing Center
550 Capitol St. NE #215
P.O. Box 2148
Salem OR 97308-2148

Re: In the Matter of PUBLIC UTILITY COMMISSION OF OREGON Investigation into Forecasting Forced Outage Rates for Electric Generating Units
Docket No. UM 1355

Dear Filing Center:

Enclosed please find an original and five (5) copies of the Reply Testimony and Exhibits of Randall J. Falkenberg on behalf of the Industrial Customers of Northwest Utilities ("ICNU") in the above-referenced docket.

We will be providing electronic copies of Randall J. Falkenberg's workpapers to the parties that have requested them.

Thank you for your assistance.

Sincerely yours,

/s/ Martin H. Patail
Martin H. Patail

Enclosures
cc: Service List

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that I have this day served the foregoing Reply Testimony and Exhibits of Randall J. Falkenberg on behalf of the Industrial Customers of Northwest Utilities upon the parties, on the official service list shown below for UM 1355, via U.S. mail to parties which have not waived paper service, and via electronic mail to the entire service list.

Dated at Portland, Oregon, this 13th day of August, 2010.

/s/ Martin H. Patail
Martin H. Patail

(W) CITIZEN'S UTILITY BOARD OF OREGON
OPUC DOCKETS
GORDON FEIGHNER
ROBERT JENKS **(C)**
G. CATRIONA MCCRACKEN **(C)**
RAYMOND MEYERS **(C)**
KEVIN ELLIOT PARKS **(C)**
610 SW BROADWAY - STE 308
PORTLAND OR 97205
dockets@oregoncub.org
gordon@oregoncub.org
bob@oregoncub.org
catriona@oregoncub.org
ray@oregoncub.org
kevin@oregoncub.org

DEPARTMENT OF JUSTICE
MICHAEL T WEIRICH **(C)**
REGULATED UTILITY & BUSINESS SECTION
1162 COURT ST NE
SALEM OR 97301-4096
michael.weirich@doj.state.or.us

(W) IDAHO POWER COMPANY
CHRISTA BEARRY
REX BLACKBURN
LISA D NORDSTROM **(C)**
GREGORY W SAID **(C)**
TIM TATUM
SCOTT WRIGHT **(C)**
PO BOX 70
BOISE ID 83707-0070
cbearry@idahopower.com
rblackburn@idahopower.com
lnordstrom@idahopower.com
gsaid@idahopower.com
ttatum@idahopower.com
swright@idahopower.com

MCDOWELL & RACKNER PC
ADAM LOWNEY
KATHERINE MCDOWELL
WENDY MCINDOO **(C)**
LISA F RACKNER **(C)**
419 SW 11TH Ave., Suite 400
PORTLAND OR 97205
adam@mcd-law.com
Katherinemcd-law.com
wendy@mcd-law.com
lisa@mcd-law.com

(W) PACIFIC POWER & LIGHT
MICHELLE R MISHOE **(C)**
825 NE MULTNOMAH STE 1800
PORTLAND OR 97232
michelle.mishoe@pacificorp.com

(W) PACIFICORP OREGON DOCKETS
OREGON DOCKETS
825 NE MULTNOMAH ST
STE 2000
PORTLAND OR 97232

oregondockets@pacificcorp.com

PORTLAND GENERAL ELECTRIC
RATES & REGULATORY AFFAIRS – OPUC FILINGS
121 SW SALMON ST 1WTC-0702
PORTLAND OR 97204
pge.opuc.filings@pgn.com

PORTLAND GENERAL ELECTRIC
DOUGLAS C TINGEY ©
121 SW SALMON 1WTC13
PORTLAND OR 97204
doug.tingey@pgn.com

PUBLIC UTILITY COMMISSION OF OREGON
KELCEY BROWN
PO Box 2148
Salem, OR 97301
Kelcey.brown@state.or.us

(W) = Waive Paper Service

BEFORE THE PUBLIC UTILITY COMMISSION

OF OREGON

UM 1355

In the Matter of	)
	)
The Public Utility Commission of Oregon	)
Investigation into Forecasting Forced Outage	)
Rates For Electric Generating Units.	)
	)

REPLY TESTIMONY OF

RANDALL J. FALKENBERG

ON BEHALF OF

THE INDUSTRIAL CUSTOMERS OF NORTHWEST UTILITIES

August 13, 2010

Q. HAVE YOU PREVIOUSLY FILED TESTIMONY IN THIS DOCKET?

A. Yes, on April 7, May 13, and August 13, 2009, I filed direct, reply, and supplemental reply testimony on behalf of the Industrial Customers of Northwest Utilities ("ICNU").

Q. WHAT IS THE PURPOSE OF THIS NEW REPLY TESTIMONY?

A. I reply to the July 16, 2010 direct testimony filed by Portland General Electric ("PGE") and Idaho Power Company ("IPC") addressing the ICNU collar mechanism proposed in ICNU/300, which was my supplemental reply testimony.

Q. PLEASE PROVIDE SOME BACKGROUND CONCERNING THIS CASE.

A. Since 1984, the Oregon Public Utility Commission (the "OPUC") has forecast generator Forced Outage Rates ("FOR")[1/] using a four-year moving average. The fundamental remaining issue is how to replace unusually high or low outage rates ("extreme outage rates" or "outliers") in the four-year average. In its prior supplemental testimony (PPL/102), PacifiCorp proposed a "collar mechanism" which would have replaced outliers with observations falling two standard deviations from the mean. In my supplemental reply testimony, I used the PacifiCorp data[2/] to develop an alternative to the PacifiCorp collar mechanism, which I call the "ICNU collar."

Q. PLEASE SUMMARIZE THE ICNU COLLAR MECHANISM.

A. The evidence demonstrates that extreme outage events are likely to be followed by "closer to normal" or average years. Therefore, extreme outages should be

---

1/ To avoid confusion, in this context I will use FOR and Equivalent FOR ("EFOR") interchangeably, as PGE did in its testimony. All data used in this analysis EFOR data.

2/ PacifiCorp provided annual outage rate data for the period 1989 to 2008 for 26 coal plants. This data has been the basis for all subsequent analysis by ICNU and PGE. There was 20 years of data for 19 of the units, and less for the remaining 7 units.

replaced with average outage rates instead of "slightly less extreme" outages recommended by PGE and PacifiCorp. The ICNU collar excludes annual FORs falling outside of the 90th and 10th percentiles ("90/10") computed from the unit history (with outages capped at a maximum length of 28 days)[3/] from the four-year moving average calculation. Those observations are replaced with the long term unit average. The analysis I presented in my supplemental testimony demonstrated that, because the FOR series is "mean reverting," this approach would improve forecast accuracy. Mean reversion implies the data will tend to return to its mean value after an unusual observation occurs. Quite simply, the question before the OPUC is whether extreme outage rates should be replaced by ones that are "closer to normal" or just "slightly less extreme."

Q. WHAT ARE THE MAIN POINTS OF THE PGE TESTIMONY?

A. PGE made the following points:

1. PGE asserts that it is makes no sense to believe the FOR series is mean reverting because the data series is "non-stationary"[4/] and best modeled by use of a four-year moving average which implies a non-stationary series.[5/]

2. PGE contends that the analysis presented in my supplemental testimony demonstrating the forecast accuracy gains from the ICNU collar suffered from a fatal error - use of ex-ante data not available at the time when forecasts would have been prepared and provides a meaningless comparison.[6/]

[3/] Re PacifiCorp, Docket No. UE 191, Order 07-446 at 21 (Oct. 17, 2007). This is a very important step in the process and was applied by ICNU, PGE and PacifiCorp in their collar analyses. This step, by itself, improved forecast accuracy by more than 9% compared with use of a simple, unadjusted four-year moving average. This improvement is highly significant under the permutation and conventional statistical test (p-value <1%).

[4/] A stationary series is one where means and variances do not change significantly over time. A non-stationary series is one where the means and variances do change significantly over time.

[5/] PGE/300, Tinker-Weitzel/5.

[6/] Id. at 1-2.

3. PGE further contends that even accepting the results of my analysis, they fail to approach statistical significance. PGE presents a series of "permutation tests" in support of this argument.[7/]

4. PGE contends that use of Root Mean Square Error ("RMSE") is not the best metric for deciding between forecast methods. PGE suggests this overstated forecast accuracy gains presented in my supplemental testimony.[8/] PGE proposes use of the Relative Geometric Root Mean Square Error ("RGRMSE") because it is allegedly less sensitive to outliers.

Q. IS THE PGE TESTIMONY CORRECT?

A. No. PGE's criticisms above are wrong, exaggerated and irrelevant:

1. **Use of a four-year moving average model <u>implies</u> a stationary (mean reverting) rather than non-stationary series. Time series analysis provides evidence the FOR series are stationary and mean reverting.**

2. **At the very most, use of ex-ante data implies it would be inappropriate to characterize the analysis in my supplemental reply testimony as a "backcast."[9/] However, it is still a very useful empirical analysis explaining the actual behavior of outage rates in the years following extreme outage rate occurrences. These analyses showed that extreme outage rates are more likely to be followed by "closer to normal" rather than "slightly less extreme" outage rates.**

3. **The forecast accuracy comparisons in my supplemental testimony are quite reasonable, and even conservative. When possible impacts stemming from use of ex-ante data was removed, the ICNU collar still provides substantial accuracy gains relative to the 90/10 replacement strategy PGE favors.**

4. **Even if the OPUC were convinced the FOR series are not mean reverting, simply excluding the outliers or replacing them with the prior year's four-year moving average provides a better forecast than the 90/10 replacement strategy.**

5. **PGE's permutation tests contain a basic error that invalidates their results. Corrected permutation tests (and conventional statistical tests) demonstrate a very high likelihood the ICNU method will improve forecast accuracy, and very low likelihood that the results obtained were due to random chance.**

7/ <u>Id.</u> at 2.
8/ <u>Id.</u> at 4.
9/ A hypothetical recreation of a forecast prepared at some time in the past.

6. **Large, but unpredictable outliers included in my supplemental testimony comparisons, understated, rather than overstated the advantage of the ICNU collar. Consequently, PGE's primary argument for RGRMSE is invalid.**

7. **PGE's criticism of the ICNU collar's accuracy comparisons are also irrelevant because even based on RGRMSE, the ICNU collar method improves forecast accuracy.**

Q. DO YOU HAVE ANY OTHER COMMENTS?

A. Yes. There are several other important points I would like to make:

1. PGE relies on two unique and novel statistical methods in this case (the RGRMSE and permutation tests). The Company has never applied either technique in any previous OPUC proceeding. It applied both techniques incorrectly, and produced misleading results.

2. PGE presents "data to date" analyses, based on use of only ex-ante data. PGE contends these analyses show the ICNU collar fails to provide forecast accuracy improvements. However, PGE's analysis contains a mathematical error and uses such a limited data sample (as few as one data point to compute the replacement mean FOR value) that the PGE results are meaningless. Instead of providing "data to date" the information used was often several years out of date.

3. Putting all other issues aside, analysis of the PacifiCorp data shows that the year following an extreme outage rate is much more likely to be "closer to normal" than just "slightly less extreme." As a result, mean replacement is a much better strategy than PGE's preferred 90/10 replacement strategy.

4. PGE presents other analyses intended to address the favorability of the 90/10 collar using NERC data as a replacement strategy.[10/] These results do not provide any insights into the forecast gains arising from a NERC based collar due to lack of data, and the other infirmities in PGE's analysis.

5. Even accepting PGE's major claims, they provide no basis for preferring the alternatives to the ICNU collar. They merely imply that a statistical analysis can't decide which alternative is best, not which is better. Even so, there are other logical or policy grounds favoring acceptance of the ICNU or OPUC collar proposals that have not been addressed by PGE.

[10/] To avoid confusion, I will refer to PGE's recommended collar (90/10 NERC data range, 90/10 replacement strategy) as the "PGE Collar." PGE refers to this as the "Staff" collar, though it is not clear the Staff still supports this method. I will refer to the collar proposed by the Commission in its Order of October 6, 2009 as the "OPUC Collar."

6. I have reservations about the way in which the NERC data has been constructed to produce a 90/10 collar. The current construction may result in a collar that is "too loose."

Mean Replacement and Omniscience

Q. WHY DOES PGE OBJECT TO THE ICNU COLLAR?

A. The ICNU collar replaces outlier[11/] observation with the long term average. PGE contends the accuracy comparison I presented in my supplemental reply testimony was "unfair" because I replaced the excluded observation with information (ex-ante data) not available when the hypothetical forecast would have been prepared. PGE states on this basis, that I was acting as if I were "omniscient"[12/] because both ex-post and ex-ante data is used in the analysis of the ICNU collar.

Q. IS THIS A VALID CRITICISM?

A. No, and although I strongly dispute this criticism, PGE greatly exaggerates the significance of this point. At the very most, use of ex-ante data would imply the analysis I presented should not be characterized as a "backcast"[13/] but rather as an empirical analysis of past behavior, much like an econometric model used for load forecasting purposes. For our purposes, there is very little difference. The analysis I performed showed that in the past twenty years the historical outage rate pattern is explained better by the assumption that after extreme outage rates occur, they are more likely to return to "normal" rather than staying fixed near the extreme levels. This provides very useful information for deciding what is the

---

11/ Based on the most recently filed testimony, 90/10 is the range supported by Staff, ICNU and PGE, while PacifiCorp supports use of a two standard deviation collar. PGE and Staff support use of NERC data for determining the collar range, while ICNU and PacifiCorp support use of plant history to determine the collar range.

12/ PGE/300, Tinker-Weitzel/7.

13/ A backcast is a hypothetical forecast done at some prior time using data available at that time. An empirical model tries to explain prior data, seeking the "best fit." Empirical models are the basis for nearly all forecasting applications.

best forecasting method from this time forward. This is the ordinary process used in building forecasting models. Econometric load forecasting models, for example, are accepted on the basis of determining the specification that best explains the historical data, not by a series of "backcasting" experiments.

Q. HAVE PGE AND PACIFICORP BEEN CONSISTANT IN AVOIDING USE OF EX-ANTE DATA?

A. No. Both PGE and PacifiCorp used ex-ante data in the analyses they presented of the collar mechanisms to establish the collar ranges (two sigma for PacifiCorp and 90/10 for PGE). PGE acknowledged its own use of ex-ante data, but asserts that this is not as serious of a "transgression" as my use of the long term average.[14/] PGE's reasoning is unpersuasive because it effectively assumes that the boundaries of the distribution are constant, even though there is no permanent mean. Under PGE's logic, the mean could actually move outside of the collar, yet the ranges would remain constant – a clearly illogical assumption.

Q. DO YOU AGREE WITH PGE THAT THE RESULTS SHOW YOUR USE OF EX-ANTE DATA INTERJECTS UNREASONABLE BIAS INTO YOUR RESULTS?

A. No. In the following, I will show that because the FOR series is mean reverting, in principle, no bias was introduced, and further, that the use of ex-ante data had a limited effect. None of my original conclusions would be changed if ex-ante data alone was used.

Q. IN APPENDIX 1, PGE CLAIMS USE OF EX-ANTE DATA WOULD ALLOW FOR PERFECT FORECAST ACCURACY. IS THIS CORRECT?

A. No. PGE argues that I *could* have changed the weights used on computing the long term average each year (1/20 for a 20 year average) to 1 for the then "current"

14/ Id. at 6.

year, and zero for all other years. However, I made no such adjustment to the weights. In fact, the "current year" observation is weighted at 5%, not 100% as assumed in Appendix 1 of PGE/300. This means that the ICNU collar assigns a weight of 95% to the 19 "incorrect" observations and only 5% to the one "correct" observation. PGE also ignores the fact that in using a four-year average, the weight of the "correct" input is typically reduced to 1/80, or 1.25%.[15/] This "contamination"[16/] has very little practical consequence for this reason.

Q. DOES THE USE OF ACTUAL "FORWARD" DATA "BIAS" THE FORECAST COMPARISON?

A. No. The "forward" data only provides a "wrong answer." For example, the actual data for 1993 is not the same as the actual data for any year before, or later (except by mere coincidence). Consequently no bias is introduced by use of any data other than the "current year" data. Further, use of forward looking data has no beneficial impact on forecast accuracy, as compared to use of prior data for FORs. A forecast based on one year or four year "ahead" data for the FOR series is not as good as a forecast based on the historical four-year moving average.[17/] If PGE were correct, then one should expect that ex-ante data does a better job of forecasting than ex-post data. This also strongly suggests that the FOR series is

---

15/ 1/20th for a 20 year data set times ¼ in the four-year moving average.

16/ The only data that could improve the forecast accuracy (thus "contaminating" the results) is the inclusion of the "current year" actual data in the long term average calculation when it replaces an outlier event. For example, if the collar excludes the 1992 observation for unit 1, then the four-year moving average for the subsequent years contains a small piece of the actual data. The 1993 forecast would contain a small piece of the 1993 actual data.

17/ A forecast of current year FORs using data one year ahead forecast produces the Sum Squared Error ("SSE") of 25.8. A forecast using a forward looking four-year moving average produces the SSE of 20.3. The traditional four-year moving average based on historical data is better, with the SSE of 18.9.

stationary, because if it weren't then use of forward data should be more effective for forecasting purposes.[18/]

Q. HAVE YOU QUANTIFIED THE IMPACT OF REMOVING THE POSSIBLE BIAS FROM THE USE OF THE CURRENT YEAR DATA FROM THE FORECAST ACCURACY COMPARISON IN ICNU/300?

A. Yes. Removing the impact of the "current year" component from the ICNU collar forecasts increases the SSE by about 2%.[19/] This is not nearly enough to change any of the conclusions presented in my supplemental testimony, as will be demonstrated later.

Q. ARE THERE OTHER REASONS WHY THE USE OF EX-ANTE DATA DOES NOT INVALIDATE YOUR RESULTS?

A. Yes. This question goes directly to the matter of whether the FOR series is stationary or non-stationary. PGE merely *asserts* the FOR series is non-stationary. PGE also asserts on the same basis, that the ICNU collar is illogical, because a non-stationary series has no permanent mean. This term that has a specific technical meaning in time series analysis. It cannot be established merely by *assertion*, but instead *requires* analysis of the data. PGE presents no such analysis.

Q. PLEASE EXPLAIN THE DIFFERENCE BETWEEN STATIONARY AND NON-STATIONARY TIME SERIES AND PROVIDE SOME EXAMPLES.

A. In a stationary series, a "shock" or unexpected change in the system (such as an extreme outage event) will sooner, or later, "dissipate" and leaves no lasting impact. In a non-stationary series, a change to the system has a long-lasting effect.

18/ PGE makes this very point in PGE/300, Tinker-Weitzel/7 using a well known example of a non-stationary series, US GDP.

19/ The SSE is the sum of the squared difference between actual and forecast data. It is important because a lower SSE means a more accurate forecast.

An example of a stationary series might be the performance of a complex machine. If a part gets fouled, (e.g., copper build up on a generator component) performance declines. However, if the part is cleaned performance reverts to prior levels.

Electricity sales are usually a good example of a non-stationary series. Whenever a customer adds a new appliance, makes an energy efficiency improvement, or whenever a new home, office or factory is built, there is a permanent change in electricity sales levels. Consequently, the real question here is whether we are dealing with a significant, permanent change to the system, or not. Economic data (e.g., consumer prices, the GDP, etc.) are also frequently examples of non-stationary series.

Q. WHY IS THIS DISTINCTION IMPORTANT?

A. PGE'S criticism would be more valid *if* the outage rate series were "non-stationary." A series is stationary if its underlying statistical properties don't change over time. If so, it does little more than fluctuate about its mean value. A series is "non-stationary" if its underlying statistical properties are changing over time. In such a case, there is no permanent mean. Such series are often unpredictable, from past data alone. Stock market prices are thought to be a classic example of a "random walk" – a non-stationary series. PGE is essentially suggesting that outage rates follow a random walk. In that case, the most recent observation, rather than a four-year moving average is more likely to be a better forecast model. This contradicts PGE's support of the four-year average.

If the outage rate data is stationary, then a long term average value computed in 1993, for example, would not be significantly different from the value computed in 2010. As a result, we can assume that in principle, the long term average computed in prior years from available data would not have been significantly different from one computed using presently available data.

Q. WOULD THE DATA NECESSARY TO COMPUTE THE LONG TERM AVERAGE HAVE BEEN AVAILABLE IN THE PAST?

A. Yes. PacifiCorp's response to ICNU Data Request ("DR") 9.1, demonstrates that 24 of its 26 coal units began operation prior to 1984 and 18 began operating in or before 1978.[20/] By 1993 (the first year used in the collar accuracy testing analyses) there should have been at least 10-15 years of data available for nearly all of the units. ICNU/402, Falkenberg/1.

There is no reason to believe that starting in 1993 (the first year used in forecast accuracy comparisons) one couldn't have computed a valid long term average or that it would differ significantly from the one computed in 2008. In this instance, PGE has confused lack of readily available data, with the non-existence of data. As the OPUC has used the four-year rolling average since 1984, there's no reason to believe that the outage rate data was never available, even if it is not readily available now. This additional data could have been used in PGE's "data to date" analyses.

20/ Both the ICNU collar and PGE's proposed collar rely upon data from PacifiCorp generating units.

Q. DID PGE ATTEMPT TO ACQUIRE ANY ADDITIONAL FORCED OUTAGE DATA FROM PACIFICORP?

A. No.[21/] In fact, PGE did not even include the additional data it already possessed for the Colstrip units, which spanned a longer time period than the data provided by PacifiCorp.[22/] However, I have been involved in PacifiCorp cases for many years, and in a prior case, I was provided outage event data for PacifiCorp generators for the period 1979 to 2004 on a non-confidential basis. This additional information will be used in subsequent analysis, to demonstrate PGE's contentions are incorrect, and that PGE's claimed "corrections" to my prior analyses provide erroneous conclusions.

Q. HAS PGE PRESENTED ANY ANALYSIS TO TEST WHETHER THE FORCED OUTAGE RATE SERIES IS STATIONARY, OR NON-STATIONARY?

A. No. The Company asserts (erroneously) that use of a four-year moving average model is inconsistent with the assumption that the outage rate series is stationary (and therefore mean reverting):

> The entire history of annual FORs for a given plant can best be viewed as a *non-stationary* time series. For our purposes, this means that the expected annual FOR- evolves over time, i.e., it is not constant. This view accords with common sense: A coal plant at the present time is the sum of repairs and upgrades that have occurred over its history; we don't expect the plant's expected FOR to be constant over time because the composition of the plant is not constant over time. This is why Parties have concluded that using recent data (the most recent four-year average) is the best way to forecast next year's FOR.[23/]

21/ ICNU/402, Falkenberg/2 (PGE Response to ICNU DR 4.4).

22/ These are the only units larger than 600 MW, and are readily identifiable from the data PacifiCorp provided PGE.

23/ PGE/300, Tinker-Weitzel/14-15.

PGE provides no evidence to support this statement. While there are factors that might degrade performance over time (plant aging), there are others (plant improvements, technology, management initiatives, etc.) that are intended to improve outage rates. A review of the data shows that neither has a permanent effect. Likewise, NERC data for the period 1997-2001 showed an average Equivalent Availability Factor for coal plants of 84.14%, as compared to 84.19% for 2005-2009 data.

Q. DOES PGE APPLY THE ABOVE "COMMON SENSE" ASSUMPTION IN OTHER SITUATIONS WHERE OUTAGE RATES ARE IMPORTANT?

A. No. PGE assumes constant unchanging outage rates for its Integrated Resource Plan ("IRP").[24/] Were PGE convinced that outage rates are changing over time, it would seem they should at least try to forecast the direction of change for IRP purposes.

Q. DID PGE ACCURATELY CHARACTERIZE THE POSITION OF PARTIES VIS-À-VIS THE FOUR-YEAR MOVING AVERAGE IN THE PASSAGE QUOTED ABOVE?

A. No. PGE mischaracterizes the position of the parties. There was actually very little analysis of this issue by the parties during this proceeding. Use of the four-year average was discussed, but there was no quantitative analysis presented during the workshop process by any party. In this case, PGE is providing a post facto explanation of only its own position. I believe the four-year average has survived this process because of longstanding tradition, and because no one has presented a "better idea."

24/ ICNU/402, Falkenberg/3-5 (Responses to ICNU DR 4.32 and 4.33).

Q. IS PGE CORRECT THAT A MOVING AVERAGE MODEL IMPLIES A NON-STATIONARY (NOT MEAN REVERTING) PROCESS?

A. No. In conventional time series analysis, a Moving Average process of order N can be represented by a model expressed as:

$$Y(t) = \mu + \sum \square(k) * \varepsilon(t-k) \quad k=1, N$$

where $\varepsilon(k) = Y(t-k) - \mu$. In this case, ε is a random error term, or residual, representing the deviation of each observation from the series mean. □(k) is the coefficient of the residual lagged k times. In such a model the forecast of the current observation is a function of the series mean and prior error terms. This model is clearly mean reverting because the forecast of future values of ε is unknown, but assumed to be zero, by specification of the model. Moving Average models imply a stationary series.[25/] The four-year rolling average model is nothing more than a special case of the above equation: N = 4, and □ (k) = ¼ for all k.[26/]

There is a companion type of process, called "Autoregressive" ("AR") because the current observation, Y(t), depends on prior observations and random error terms. Generally, such processes are also stationary, and like the Moving Average the observations tend to fluctuate around a fixed mean.[27/]

The proper model for a non-stationary series is typically called a "random walk" model, and implies that the most recent data point is the best forecast of future observations. Exponential smoothing models are also used for this situation. Neither is comparable to a four-year moving average.

25/ George P. Box, and Gwilym M. Jenkins, Time Series Analysis: Forecasting and Control 67 (rev. ed. 1976).

26/ $Y(t) = \mu + 1/4(\varepsilon(t-1) + \varepsilon(t-2) + \varepsilon(t-3) + \varepsilon(t-4))$; $Y(t) = ¼((\mu + \varepsilon(t-1)) + (\mu + \varepsilon(t-2)) + (\mu + \varepsilon(t-3)) + (\mu + \varepsilon(t-4)))$. Based on the definition of Y(k), and $\varepsilon(k)$: $(\varepsilon(k) = Y(k) - \mu)$. $Y(t) = ¼(Y(t-1)+Y(t-2)+Y(t-3)+Y(t-4)$ which is the conventional formulation.

27/ George C. Tiao, An Introduction to Applied Time Series Analysis (1975); Box and Jenkins, supra, at 56.

There is <u>no reason</u> to assume that simply because a moving average model is used to represent a series that it implies that the series is non-stationary, and not mean reverting. In fact, just the opposite is true, based on traditional time series modeling techniques. A careful reading of the response to ICNU DR 5.34, shows that PGE admits that when the order of moving average process is finite (e.g., 4), the series <u>is</u> stationary. ICNU/402, Falkenberg/6-7 (PGE response to ICNU DR 5.34). This shows PGE now admits that a fundamental premise of its criticism of the ICNU collar is incorrect.

In fact, PGE's entire line of reasoning on this point is backwards. Simply *assuming* that a moving average model is appropriate implies *nothing* about the underlying series or actual data. There are various statistical tests available to determine whether a moving average model is appropriate and whether a time series is stationary, or non-stationary. PGE references these tests in the response to ICNU DR 5.34, but did not present any such analysis in its testimony.

Q. HAVE YOU PERFORMED A TIME SERIES ANALYSIS TO DETERMINE WHETHER THE FORCED OUTAGE RATE SERIES ARE STATIONARY?

A. Yes. I used a variety of accepted techniques and statistical packages, described in Exhibit ICNU/401. The analysis of monthly, semi-annual and annual data provides strong evidence the FOR series are stationary, and mean reverting. This invalidates PGE's first two arguments.[28/] That being the case, the results presented in my supplemental testimony provide a fair and meaningful test of the ICNU collar proposal. Further, this demonstrates the theoretical validity of the ICNU collar mechanism because stationary series are mean reverting as shown already

[28/] PGE arguments were that: 1) It was illogical to assume the FOR series was mean reverting; and 2) Using the long term mean data produced an unfair bias in the forecast accuracy comparisons.

using simpler methods in my supplemental testimony. ICNU/300, Falkenberg/7. PGE did not address that analysis.

Q. DOES THIS MEAN THAT THERE ARE NEVER ANY TRENDS APPARENT IN OUTAGE RATE DATA?

A. No. A stationary series can exhibit short term trends because there can be a relationship between current and prior observations. However, eventually these series will revert back to average values. Concerns about the apparent trends in the outage rate data may have been part of the impetus for this investigation. While short term trends have appeared in the past, over the long term, there is no evidence of a permanent trend in the FOR series. Consequently, a mean reversion model is both logical and reasonable.[29/]

Q. ASSUMING THE FORCED OUTAGE RATE SERIES ARE STATIONARY, DOES THE FOUR-YEAR MOVING AVERAGE PROVIDE AN ACCEPTABLE STARTING POINT FOR FORECASTING PURPOSES?

A. Yes, though, use of the four-year moving average would likely not have been the result of a rigorous time series analysis.[30/] However, I have examined the issue and believe the four-year moving average model is still acceptable. This analysis also has an important bearing on PGE's other contentions.

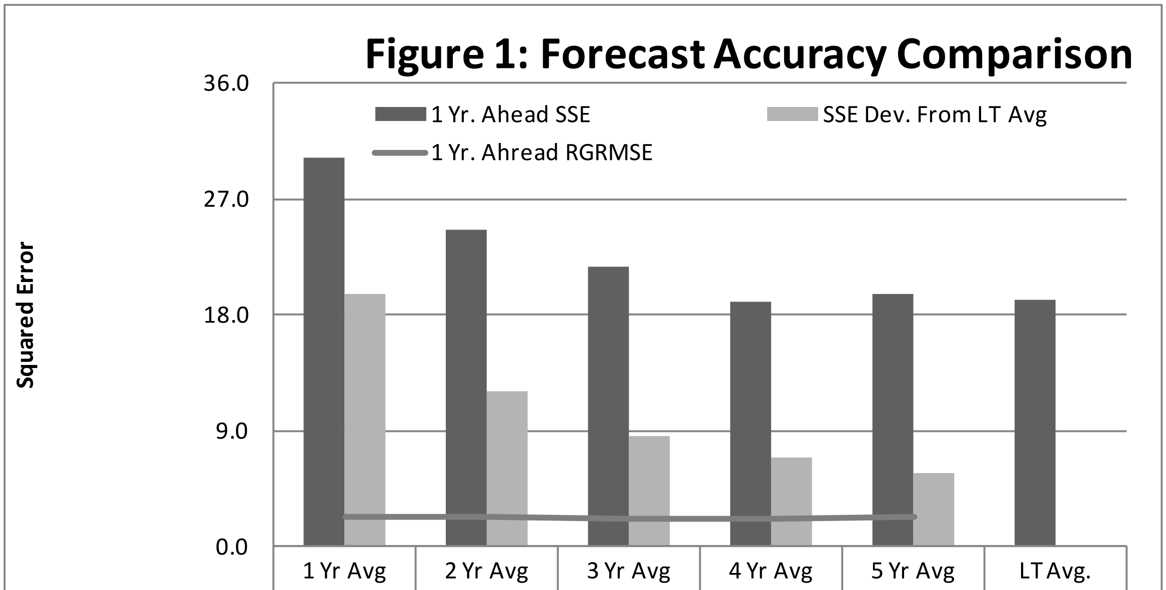
I have analyzed this issue by comparing year ahead forecasts of the results based on one through five year moving averages along with an analysis of the variances between the long term average and the moving averages. The figure

---

29/ The only instances where outage rates are likely to vary over time is in the first year or two of operation and the final few years. Once the outage rate "matures" it is likely to be flat, until the end of a unit's lifetime. At the very end of a unit's life, outage rates can decline as it is uneconomic to make further investment in the plant. This "bathtub curve" is the conventional assumption for outage rate modeling in planning studies. It is the period between the initial, and final years of operation we are dealing with in this discussion.

30/ Indeed, the "signature" of a MA(4) model was generally not present in the time series analysis I performed of the FOR series.

below shows the results of this analysis. It also addresses the question of whether the use of the long term average biased the forecast accuracy comparisons I presented in my supplemental testimony.



	1 Yr Avg	2 Yr Avg	3 Yr Avg	4 Yr Avg	5 Yr Avg	LT Avg.
1 Yr. Ahead SSE	30.1	24.5	21.7	18.9	19.6	19.2
SSE Dev. From LT Avg	19.5	12.0	8.6	6.9	5.8	
1 Yr. Ahread RGRMSE	2.24	2.37	2.14	2.19	2.32	

Q. PLEASE EXPLAIN THIS CHART.

A. The dark bars on the chart compare the average Sum Squared Error ("SSE") for various forecast models using one through five year moving averages, based solely on ex-post data.[31/] Use of a one year average (i.e., using the single prior year) to forecast the subsequent year produces an average SSE of 30.1, while a two year moving average lowers the average SSE to 24.5. Use of a 4 year moving average produces an average SSE of 18.9. Though the four-year moving average produces the lowest sum-squared error, there is little difference between the results for four years or longer periods. Indeed, the last column (labeled LT Avg) represents the

31/ I will discuss the line in the chart above in the discussion of RGRMSE.

average for the entire available data period.[32/] The long term average produces a nearly the same result (18.9 v. 19.2) as the four-year moving average. Since the long term average result differs little from the four-year moving average, I believe this also demonstrates that no meaningful bias was introduced by using the long term average instead of the excluded data. Indeed, the chart above shows that using the long term average carte-blanche is marginally *less efficient* than the ex-post four-year average. If PGE's comment about biasing the result by use of ex-ante data were valid, then one would expect the long term average figure to outperform the moving averages based solely on ex-post data. Later I will present results of other analyses which use only ex-post data in the ICNU collar.

Q. DOES THIS CHART ALSO HELP EXPLAIN WHY THE FOUR-YEAR AVERAGE PROVIDES A PRACTICAL FORECASTING TOOL?

A. Yes. The chart also shows that the use of a four-year moving average captures the mean reverting tendency of the data better than a shorter series, though a longer term average does a slightly better job. The second bar (the light gray columns) shows the departure of the moving averages from the long term average. The data shows that increasing the order of the moving average model reduced the deviations from the long term average. Quite simply a four-year moving average forecasts the long term average better than a three moving average due to mean reversion. Using only the most recent data (one year) produces a very poor forecast (having a sum-squared error of 30.1) because it fails to replicate the mean reverting tendency of the data (differing from the long term average by 19.5). The data shows that a four-year moving average departs by far less from the long term

32/ Twenty years for most plants, nine for others.

average. This data provides a practical basis for continued use of the four-year moving average, though it the lacks statistical rigor of a true time-series analysis. The alternative would be a unique, custom built time series model for each generator, which seems impractical, and beyond the scope of this round of testimony.

Q. IF THE FORCED OUTAGE RATE SERIES IS STATIONARY, WHY NOT SIMPLY USE THE LONG TERM AVERAGE TO FORECAST OUTAGE RATES?

A. That is a plausible and practical solution. On a purely statistical basis there is little reason to do otherwise. Use of the long term average could largely eliminate the necessity of deciding the issue of replacement strategies for outliers. The problem with a four-year moving average is that a "one in twenty" event is given far too much weight. If we were using a twenty year average, a single observation would arguably not distort the forecast at all, particularly is "extreme events", longer than 28 days were capped per OPUC Order 07-446.[33/]

However, there is a "downside" to using the longer data period – it would require more data and more analysis. This is one reason to prefer shorter data periods, as processing the data in a rate case setting is time consuming and schedules are short. In any case, the chart above shows that little accuracy in forecasting is gained or lost once we use a moving average of four years or more. Simplicity and tradition would favor retaining the status quo. However, if a four-year average is used, we do need to decide how to deal with outliers in the best possible manner. PGE endorses a replacement strategy based on exchanging extreme events for slightly less extreme events (the 90/10 replacement strategy). I

[33/] Re PacifiCorp, Docket No. UE 191, Order 07-446 at 21 (Oct. 17, 2007).

disagree. In the next section of the testimony, this issue will be addressed and PGE's criticisms of the ICNU collar will be addressed.

Permutation Tests

Q. THE THIRD LEG OF PGE'S CRITICISM OF THE ICNU COLLAR IS THAT THE FORECAST IMPROVEMENTS LACKED STATISTICAL SIGNIFICANCE. IS PGE CORRECT?

A. No, but even if true, it doesn't mean the PGE (or any alterative) collar is "better." It simply means that statistics can't tell which is best. In that case, logic, common sense, and good policy should shape the decision. On those grounds, the ICNU collar is superior and should be adopted.

Further, PGE bases its argument on use of "permutation tests." The PGE results, suggest a high likelihood that the forecast accuracy improvements reported in my supplemental testimony were simply coincidental. However, PGE made a very basic error in these tests which, once corrected, demonstrates that the accuracy improvements are not coincidental.

Q. PLEASE EXPLAIN.

A. PGE's analyses are wrong because the permutation tests were performed incorrectly. PGE used a testing method appropriate for two randomly drawn samples, rather than using a testing method appropriate for a single sample of "matched pairs." This mistake foreordained the results of the permutation test to show a lack statistical significance. PGE's mistake was to exchange, non-exchangeable data – to confuse a "before and after" sample with two random samples.

Q. PLEASE EXPLAIN.

A. PGE's permutation tests are only appropriate for a comparison of two randomly selected samples. An example might be a drug trial, where patients are put into either a control or test sample. To test a new drug treatment a common approach would be to select 2000 patients at random, give 1,000 the trial drug and 1,000 a placebo. Assuming the patients could be scored on severity of symptoms from 1 to 5 the scores of the two groups could be compared. If the control group had an average score of 3 while the group that received the drug had a score of 2.5, the drug shows an average improvement of .5. However, it is possible that the composition of the sample and variability of the scores were such that the results could were merely due to chance. For example, it is possible the trial group started with a lower average symptom score than the control group merely by chance.

There are a number of techniques used to test the significance of such results. If the sample scores were normally distributed the standard t-test could be used. However, if the samples weren't normally distributed then the t-test might not provide accurate results. The permutation test is intended to address such situation.[34/] PGE apparently used the permutation test for this reason, as it showed the sample was not normally distributed.[35/]

Q. HOW WOULD ONE PERFORM A PERMUTATION TEST FOR THE RANDOM SAMPLE DRUG TRIAL?

A. In this example, the null hypothesis is that there is no difference between the average symptom score of the two groups. If true, then it would make no

---

34/ Another application of the permutation test is for small samples, where there is not enough data to use more conventional techniques.

35/ PGE/300, Tinker-Weitzel/10.

difference which group a patient was in (the control or trial group) and moving patient scores between the groups at random should not change the results. In the permutation test one combines all the scores into one sample (i.e., put all the test results into a bowl), and then randomly selects two samples. Average scores would then be computed for each "new" sample. After performing this process hundreds of times, we could see how many times a difference between the two samples greater than or equal to .5 occurred. If only a small percentage of such randomly drawn scores resulted in an average difference of .5, then the result from the original grouping was unlikely to have occurred merely by chance. If only 5% (called the *p-value*) of the random samples produced a score result greater than .5, then the conclusion is that the new drug is effective is viewed to be statistically significant at the 5% level.

Q. IS THIS HOW PGE PERFORMED THE PERMUTATION TESTS?

A. Yes. PGE treated the various scenarios as two independent random samples and performed the test above. When I replicated their analysis based on this design, I reproduced both the p-value and the shape of histogram shown in Figure 1 of PGE/300.

Q. WHY IS PGE'S PERMUTATION TEST WRONG?

A. PGE ignored the most basic fact of all: we are not comparing two random samples, but rather one sample of "matched pairs." We started with one sample of annual unit data and from this we produced a set of matched pairs of forecast results (one for each method used). Comparison of the ICNU collar to the "Do Nothing" case (where we make no collar adjustment) could best be thought of as a "before and after" drug trial applied to the same patient (i.e., the forecast outage rate for unit x, for year y). The comparison of the ICNU collar to alternative methods could be thought of as comparing results for symptom scores of the same patient with different drugs. Use of a matched pair study design greatly increases the "power" of a statistical test, and is a useful technique for that reason.[36/] This is also why identical twins are often popular test subjects.

This is a fundamentally *different* study design than PGE assumed. The assumption of two randomly drawn samples PGE used was incorrect from the very start. In effect, PGE threw away much of the useful information in the test by breaking up the matched pairs. This was clearly wrong because the individual observations for a specific unit were highly correlated across the various forecasting strategies.[37/] In effect, PGE tested whether a procedure that compared results of one forecast strategy of Colstrip 3 in 2001, with a (possibly different) forecast strategy for Cholla 4 from 1993, would produce random results. PGE did not test whether the forecast for each unit for each year was different between the two methods considered. It should be of little or no surprise that PGE obtained random results in this kind of analysis.

Q. WHAT IS THE PROPER WAY TO PERFORM A PERMUTATION TEST FOR MATCHED PAIRS?

A. In this case, the question is whether the procedure, as applied to a specific unit in the sample improves the forecast or not. If the forecast is not improved, then it

36/ Increasing power enables use of a smaller sample size. George W. Snedecor and William G. Cochran, Statistical Methods 107 (6th ed. 1967).

37/ The scores under the various collar methods were highly correlated. A permutation test between the matched pairs showed that there was almost zero chance that the correlation between the scores of the ICNU and alternative collars occurred by chance. This demonstrates the inapplicability of PGE's permutation test methodology.

wouldn't matter whether the unit in question was part of the ICNU collar sample, or the alternative strategy sample. If so, then randomly assigning the result between the two groups, or simply randomly reversing the sign of the difference between the ICNU collar and alternative strategy results at random would produce equally valid results.[38/]

Q. CAN YOU DEMONSTRATE THAT PGE'S APPLICATION OF THE PERMUTATION TEST IS WRONG?

A. Yes. Consider an example where you have sample of matched pairs, x and y, but y =x+1. In this sample x is randomly drawn, with four observations (1, 4, 8, and 20). The resulting y values are 2, 5, 9 and 21. The difference between the two groups of measurements is four. While the difference between x and y is small, the variation among the x and y sets is large. In PGE's permutation test, the variations in x and y, overwhelm the differences between the individual x and y values.[39/] When applied to this data, the PGE permutation test shows that a difference of 4 or more can occur more than 40% of the time based on 500 simulations. However, the proper matched pair permutation test shows a difference of 4 only occurred 4% of the time in the 500 simulations. As a result, the PGE test can't draw any conclusions from the data, while the proper permutation test draws the correct conclusion. This is simply because PGE sacrificed the validity of the results and greatly reduced the power of the data, by

---

38/ In fact, this approach goes back to the Fisher Randomization test. Snedecor and Cochran, supra, at 133.

39/ This is also true in the SSE series – the difference between individual observations is often much larger than the difference between the various collar mechanism results for individual observations. In fact, many are zero, because the collar mechanism doesn't always apply.

throwing away the most valuable information in its permutation tests – the impact of the various strategies on each unit specific forecast for each year.

Table 1
Corrected Permuation Test Results p-Values

	ICNU Collar	Alternative Strategy	Accuracy % Gain	Matched Pair Permutation	Conventional	PGE	Metric
1	Standard	Do Nothing	11.5%	0.4%	0.5%	-----------	SSE
2	Standard	90/10 Replacement	5.5%	3.7%	5.9%	41%	SSE
3	No "Contamination"	90/10 Replacement	4.1%	10.2%	12.5%	-----------	SSE
4	1979-Data to Date	90/10 Replacement	3.0%	18.2%	20.0%	-----------	SSE
5	RGRMSE Optimized	90/10 Replacement	8.5%	7.0%	-----------	-----------	RGRMSE
Scenarios Capping Outages Durations at 28 days per Order 07-0446:							
6	Standard	90/10 Replacement	9.1%	0%	0.2%	-----------	SSE
7	No "Contamination"	90/10 Replacement	7.4%	1.1%	1.1%	-----------	SSE
8	1979-Data to Date	90/10 Replacement	5.9%	2.1%	1.8%	-----------	SSE
9	Prior 4 YR Avg	90/10 Replacement	2.7%	4.8%	5.1%	-----------	SSE
10	Remove Outlier	90/10 Replacement	5.0%	3.9%	4.8%	-----------	SSE

Q. PLEASE EXPLAIN THE RESULTS SHOWN ON TABLE 1.

A. Table 1 shows the p-value results for permutation tests I performed. This table also provides a correct version of a test PGE performed.[40/] The table compares various strategies using the ICNU collar mechanism as compared to the "Do Nothing" and 90/10 ("own unit data") strategies. The Table also shows the percent gain in accuracy of the ICNU collar compared to the alternative. Scenario 1 compares the ICNU collar to the "Do Nothing" strategy (i.e., making no adjustments to the data other than capping outage durations at 28 days). The p-value estimates the probability (based on the matched pair permutation test) that the advantage of this strategy occurred due to random chance. The very low p-value (.4%) indicates it is extremely unlikely that the observed advantage occurred

---

40/ In the workpapers, I provide a number of source documents that support this approach to permutation testing for matched pairs.

merely due to chance. PGE did not provide any permutation test results for this scenario.

Scenario 2 compares that ICNU collar to one using a 90/10 replacement strategy. PGE presented this strategy on Table 1 in PGE/300 reporting a p-value of 41%. This was a primary element of PGE's criticism of the ICNU collar. However, the correct p-value result is only 3.7%.

Q. EXPLAIN WHAT IS MEANT BY THE "CONVENTIONAL" TEST RESULTS.

A. A permutation test is not necessarily required for a matched pair data set. Because we are dealing with matched pairs of observations, the variable of interest, D, is the <u>difference</u> between the Squared Error ("SE") results for the ICNU and alternative collar, i units and j years:

$$D_{ij} = SE\ (ICNU)_{ij} - SE_{ij}\ (Alternative)) \equiv SEI_{ij} - SEA_{ij}$$

$$Var\ (D) = Var\ (SEI) - 2Covar\ (SEI, SEA) + Var\ (SEA)$$

The Central Limit Theorem holds that the sample mean, $D\mu = (1/N\sum Dij$ for all i j), has a variance approximately equal to Var (D)/(N) where N is the sample size (irrespective of the distributions of the Squared Error ICNU ("SEI") or Squared Error Alternative ("SEA")). Owing to a very large sample (N=354) the variance of the sample mean is small. The ratio $D\mu/\sqrt{}$ (Var (D)/N) measures the distance of the sample mean, Dμ, from zero in standard deviations. The probability distribution of the sample mean is approximately normal (again owing to the Central Limit Theorem). It is a simple matter to determine the p-value from these results using standard normal distribution tables.[41] The results above are in excellent agreement with the ICNU matched pair permutation test, but differ dramatically from the PGE results. The close agreement of these two independent tests, and the substantial deviation apparent in the PGE results, should provide ample evidence that PGE's analysis is simply wrong. I present these conventional results for permutation tests I performed where possible.

Q. PLEASE EXPLAIN THE SCENARIO 3 ON TABLE 1.

A. One of PGE's primary objections to the ICNU collar mechanism was that the accuracy comparisons presented in my supplemental testimony relied on ex-ante data. While I dispute the validity of this criticism, as discussed above, I did an analysis which used a replacement strategy which removed any possible "contamination" from inclusion of the current year actual data in the collared forecast. This raised the SSE score of the ICNU collar shown in my supplemental testimony by less than 2%.[42] The comparison I performed showed that this strategy provides a more accurate forecasting methodology than the 90/10 replacement strategy (SSE of 5472 vs. 5707). Table 1, above, shows that it unlikely (odds of one in ten) that this advantage simply occurred due to random chance.

---

[41] This technique was explained in more detail in ICNU/100, Falkebnerg/48-49. It can also be found in standard texts on statistics or econometrics. Also, the Central Limit Theorem says the distribution of the sample mean will approximate a normal distribution, with the approximation getting better as sample size increases. For this reason, for small samples, the Permutation test could be more exact. However, it needs to be performed using Monte Carlo simulations. This method is discussed in standard texts. Ralph E. Beals, Statistics for Economists: An Introduction 198 (1972).

[42] 5391 to 5472.

Q. PLEASE DESCRIBE SCENARIO 4 ON TABLE 1.

A. PGE presented scenarios showing results based solely on use of ex-post data (referred to as the "data to date" analysis). The Company claimed these results showed no accuracy gain was produced by the ICNU collar. However, PGE used a very limited data set (the original 20 year sample provided by PacifiCorp). As a result, for some of the early years in the study, the long term average used by PGE was based on only one or a few data points. As a result, PGE's approach didn't provide a fair comparison. As discussed above, I have outage rate data starting from 1979 for PacifiCorp plants. When the additional data is included, I was able to perform a much better analysis using only "data to date" starting from 1979. This provides a limited set of data which has not been fully vetted by PacifiCorp, and which I would prefer to have had more time to validate. Nonetheless, for purposes of computing a long term average based only on ex-post data, it provides a much better basis for comparison than the PGE analysis. Any inaccuracies in the data would likely diminish the results. The results indicate the accuracy gain from the ICNU collar (based solely on the limited set of ex-post data) were unlikely (odds of less than 4 to 1) to have occurred merely by chance. Though not "statistically significant" in the conventional sense it does demonstrate the likely advantage of the ICNU collar in a pure backcasting test.

Q. ASIDE FROM THE PROBLEM OF USING VERY LIMITED DATA, WERE THERE ADDITIONAL SERIOUS MISTAKES IN PGE'S "DATA TO DATE" ANALYSES?

A. Yes. PGE failed to consider that as more data was being added each year, the new information should have been included in the "replacement means" computed in the "data to date" analysis for subsequent years. PGE's calculations are only correct for the first two years that a data point is used in the four-year moving average calculation. PGE should have "updated" the replacement mean each subsequent year, but did not.

For example, Unit 17 has an outage rate that exceeded the 90th percentile in 1992. PGE computed the 1992 "replacement mean" based on 1989-1991 data (3 years). This 1989-1991 "replacement mean" for the 1992 outage rate was then used to compute four-year moving averages for the years 1992-1995. Arguably, the 1992 result was correct under PGE's assumptions. However, in the following year, additional data did become available, but PGE continued to use the same replacement mean based on 1989-1991 data. The four-year average for the last year (1995) was used for the 1996 forecast. However, by 1995 several more years of data would have been available. As a result, the 1996 forecast was based on using only three data points available as of 1991. It was fully four years out of date by that time! PGE ended up excluding as much ex-post data in that instance as it actually used. This is not "data to date" but rather a "data out of date" analysis. This same type of mistake occurred in all of the cases where data was removed because it fell outside of the 90/10 collar.

In another example, the outage rate for Unit 4 in 1991 was outside the collar. PGE used the average of the 1989 and 1990 data to compute the replacement value in the four-year averages computed for 1993 and 1994. These were then used in the 1994 and 1995 forecasts. As a result, the 1995 forecast was based on two data points available at the end of 1990. PGE ignored all of the actual data available in subsequent years.

Q. ARE THESE THE ONLY EXAMPLES OF THIS PROBLEM?

A. No. This problem occurred whenever an actual value was replaced by the collar mechanism. Consequently, PGE did not use all of the data available as of the time a forecast would have been prepared. Because of the limited data set PGE started with in the first place, and the errors in its permutation tests, the "data to date" results are of no value.

Q. CAN YOU SUMMARIZE THIS PORTION OF YOUR TESTIMONY?

A. PGE's permutation tests are seriously flawed. When a correct permutation test is applied, the results indicate it his extremely unlikely that the forecast accuracy gains of the ICNU collar compared to the "Do Nothing" alternative occurred by mere chance. It is very unlikely that the advantage of the ICNU collar as compared to the 90/10 replacement collar occurred by chance. It is unlikely that the advantage of the ICNU collar, when the impact of use of ex-ante data is removed compared to the 90/10 replacement collar (which was not limited to ex-post data), is due to chance. These analyses clearly demonstrate that PGE's third major point, as discussed above, is without merit.

Relative Geometric Root Mean Square Error ("RGRMSE") and Outliers

Q. PGE PROPOSES THAT THE OPUC RELY UPON RGRMSE TO MEASURE FORECAST IMPROVEMENT AS COMPARED TO THE SSE METRIC. PLEASE EXPLAIN THE DIFFERENCE BETWEEN THE TWO METRICS.

A. Both metrics start with the same basic data point – the squared difference between the forecast and actual outage event (again, SE for Squared Error). The SSE metric simply sums these squared errors for all observations (16 years of forecasts for 26 units). The RGRMSE metric computes the product of the squared errors for each of the N observations then takes $(2N)^{th}$ root.[43/] This process reduces the sensitivity of the analysis to inputs with high variability.[44/] The interpretation of RGRMSE is the average percentage improvement in forecast accuracy for a single observation according to the articles PGE provided. For permutation tests, PGE transformed the metric again, using its natural log.

Q. WHAT DOES USE OF THE RGRMSE DO TO THE FORECAST ACCURACY COMPARISONS?

A. RGRMSE greatly reduces the sensitivity of the comparisons – quite simply, it "flattens out the data." (Refer again to Figure 1.) The figure showed a pronounced advantage of the four-year average forecast methodology as compared to a one year average – SSE of 18.9 vs. 30.1 However, under the RGRMSE (which is also shown as the line on Figure 1) there is only a 2% difference between the two scores (2.19 vs. 2.24). Under RGRMSE, there is essentially no difference in the results of the two forecast methods. Strangely, RGRMSE would support use of a three year moving average forecast, though the advantage, again, compared to the four-year moving average is only 2%.

Q. DOES THE ICNU COLLAR STILL SHOW AN ADVANTAGE UNDER THE RGRMSE METRIC?

A. Yes. Compared to the 90/10 replacement strategy, the ICNU collar still produces an average forecast accuracy gain of 3% (1.72 for the ICNU collar vs. 1.77 for the 90/10 replacement collar).

43/ The square root for one data point, the 4th root for two data points, and so on.

44/ J. Scott Armstrong and Fred Collopy, Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons, 8 International Journal of Forecasting 69-80, 74 (1992).

Q. DID PGE FAIRLY PORTRAY THE IMPLICATIONS OF RELIANCE ON THE RGRMSE METRIC IN THE ANALYSES IT PRESENTED?

A. No. PGE may have created a misimpression that there is little difference between collar that replaced outliers with "closer to normal" observations and one using the 90/10 (or "slightly less extreme") replacement strategy. However, PGE failed to recognize that under RGRMSE, a "closer to normal" replacement strategy is still far more effective than a "slightly less extreme" replacement strategy. Scenario 5 on Table 1, addresses this issue. It demonstrates that a replacement strategy, designed to minimize RGRMSE would still replace outliers with observations closer to the middle rather than the fringes of the outage rate distribution.

Because RGRMSE is intended to help select the best forecast method, the 90/10 replacement strategy should be compared to a strategy that is optimized for RGRMSE. Based on RGRMSE the most efficient mechanism would be a 54/46 replacement strategy. Very high outliers should be replaced by the 54th percentile observations and very low ones with the 46^{th} percentile. This strategy still replaces extreme outage rates with ones "closer to normal" rather than the one that is "slightly less extreme" (90/10). The 54/46 replacement strategy minimizes RGRMSE, but not SSE. It would be the best replacement strategy if the OPUC decides that RGRMSE is the proper metric.

In the scenarios in question, the RGRMSE for the 54/46 replacement strategy is 1.63 vs. 1.77 for the 90/10 replacement collar. This is an advantage of about 8.5%. The proper interpretation of these results is that best RGRMSE based "closer to normal" replacement strategy is nearly 8.5% more accurate than the 90/10 ("slightly less extreme") replacement strategy. The permutation test results shown do indicate that it is rather unlikely that the advantage for the 54/46 replacement collar computed using the RGRMSE metric occurred only due to chance. Indeed, the p-value is only 7% implying a low probability the results occurred merely by chance.

Q. PGE ARGUES THAT YOUR RELIANCE ON THE CONVENTIONAL SSE METRIC SKEWED YOUR RESULTS IN FAVOR OF THE ICNU COLLAR. DO YOU AGREE?

A. No. PGE argues that RGRMSE is preferable because it is less sensitive to outliers than SSE. As shown above, RGRMSE is certainly much less sensitive. In PGE's view, this suggests my results overstate the benefit of the ICNU collar. While outliers do increase the SSE in absolute terms, this doesn't mean that the results from my supplemental testimony were exaggerated. In contrast, the inclusion of large outliers that the collar methods were never intended to forecast actually causes the SSE scores to *understate* the accuracy gains afforded by the ICNU collar.

Q. PLEASE EXPLAIN.

A. Some of the largest errors in the FOR forecasts were those created by "extreme events" that neither the ICNU collar, nor any other collar make any attempt to predict. These are the initial observations of outage events lasting longer than 28 days that PacifiCorp capped at 28 days in its calculation of the four-year moving average, in accordance with OPUC precedent.[45/] However the forecast error, SSE, was computed by comparing the actual (uncapped) data to the forecast values. Consequently, the individual SE observations include large outliers created when

---

[45/] Re PacifiCorp, Docket No. UE 191, Order 07-446 at 21 (Oct. 17, 2007).

the four-year rolling averages (which essentially ignored those events) were compared to actual outage rates that included these events.

As a result, this process introduced large errors into both the ICNU and alternative collar mechanism forecasts, which were never intended to be predicted by any collar mechanism. Indeed, the purpose of the collar mechanism is <u>not</u> to predict the next extended Hunter or Boardman outage, but rather how to compute outage rates subsequent to those events when circumstances "return to normal."

Q. SHOULD THE EXTREME EVENTS, SUCH AS THE BOARDMAN OR HUNTER OUTAGES BE REMOVED FROM THE FORECAST ACCURACY COMPARISONS AS WELL AS THE FORECASTS USED FOR NPC PURPOSES?

A. Yes. ICNU and PacifiCorp agree with the OPUC precedent requiring such events should be capped in outage rate forecasts. Consequently, it is more reasonable to compare the outage rates with extreme events capped to the collar mechanism forecasts. When this is done, the superiority of the ICNU mechanism is substantially greater than portrayed in my supplemental testimony. Table 2 below shows the SSE results for the ICNU collar and other scenarios based on comparison to actual uncapped, and the capped FOR data. The Table demonstrates that the original ICNU analysis was actually quite *conservative* in the presentation of the result because of the way in which these large unavoidable residuals were counted.

Table 2: Impact of Extreme Event Residuals On ICNU and Alternative Collar Results

Original ICNU/300 Page 11 Results

Strategy	Improvement (Over Do Nothing)		
	SSE	Absolute	%
Do Nothing	6,091	-	-
2 Sigma	6,003	88	1.4%
90/10 Collar	5,707	384	6.3%
ICNU Collar	5,391	699	11.5%

Results with Extreme Events Capped at 28 Days Per Order 07-446

Strategy	Improvement (Over Do Nothing)		
	SSE	Absolute	%
Do Nothing	4,546	-	-
2 Sigma	4,442	104	2.3%
90/10 Collar	4,111	435	9.6%
ICNU Collar	3,737	810	17.8%

Scale/Units: 1 represents one percent, eg. 2.3% represented by 2.3
Number of Data = 354

Q. EXPLAIN THE RELATIONSHIP BETWEEN EXTREME EVENTS AND SCENARIOS 6-10 IN TABLE 1, ABOVE.

A. Table 1 completes the analysis by including two scenarios that have extreme events capped at 28 days. Scenario 6 compares the ICNU collar to the 90/10 replacement collar. These results show that the ICNU collar clearly understated the forecast accuracy gains. Table 2 shows that this scenario produces a SSE of 3,737 compared to 4,111 to the 90/10 (the "Modified Staff") replacement collar. Table 1 shows that this result (a 9% improvement) is nearly impossible (<.2%) to be the result of random chance. In fact, it never occurred in 1000 simulations in the permutation test.

Scenario 7 removed the possible "contamination" from use of ex-ante data in the collared forecasts from the analysis by removing the actual value from the calculation of the long term average. The results show that the accuracy gain from this scenario as compared to the 90/10 replacement strategy (7.4%) has a p-value of only 1.1%.

Scenario 8 compares the "data to date" forecast strategy for the ICNU collar, based on the ex-post 1979 forward data, as compared to the 90/10 forecast. When extreme events are removed, the ICNU collar produces an accuracy gain of 6%. Scenario 8 on Table 1 demonstrates that it is highly unlikely (<2.1%) that this accuracy gain was due to random chance.

Q. WHAT IS THE PURPOSE OF SCENARIOS 9 AND 10?

A. These scenarios address the issue of replacement strategy and PGE's "data to date" scenarios. PGE argues that the FOR series is non-stationary, thus mean reversion is an incorrect assumption. If true, then use of the long term average may not be the best strategy. In Scenario 9, I compare the use of a replacement strategy based on using the prior four-year moving average. For example, if the 1993 FOR is outside of the collar, it would be replaced by a four-year moving average from 1989-1992. In scenario 10, an even simpler strategy is used: the outliers are excluded from the computation of the moving average. This will result in computing moving averages that are based on less than four data points for excluded data. In both cases, there are forecast accuracy gains relative to the 90/10 replacement strategy, and these gains have p-values of 5% or less under the permutation test. The better strategy appears to be the removal of outliers. Consequently, even if the OPUC is persuaded that the FOR series are not stationary, simply excluding the outliers provides a better forecast than using the 90/10 strategy.

Empirical Analysis of Replacement Strategy

Q. PLEASE EXPLAIN TABLE 3.

A. The most important question here is not whether the FOR series is stationary or whether use of ex-ante data produced biased results. Instead, the question quite simply is this: What is the best strategy for replacement of outliers? Will "extreme" results foreshadow a transition to a more extreme era, "a new normal," or will things return to the "old normal?" In other words, should a "closer to normal" strategy be followed or should a "slightly less extreme" replacement strategy be used?

Table 3 compares the SSE results based on using the original PacifiCorp 1989-2008 FOR data. It compares the long term average, and 90/10 replacement strategy results for the years following the best and worst years (i.e., the one in twenty events). Irrespective of whether the series is stationary or "mean reverting," this test will show what is most likely to happen after an extreme event year occurs, and which pattern exists in the actual data.

The results indicate that the year after a "one-in-twenty" event year is much closer to a normal year, than to a (slightly less extreme) "one-in-ten" year. Consequently, an extreme event is far more likely to be followed by "closer to normal" year than a "slightly less extreme" year. The table also shows that the advantage of the mean replacement strategy is not due to random chance. P-value for these results are 2.6% under the permutation test and 3.7% under the conventional test. This test itself makes much of the PGE testimony irrelevant, because it shows that "closer to normal" replacement strategy is the actual pattern in the past, and there is no reason to believe for forecasting purposes anything else should be assumed.

Table 3
Test of Replacement Strategy

		SSE
Replace Outlier with Mean		662.88
Replace Outlier with 90/10		1105.02
Advantage		442.14
Percent		40%
p-value	*Permutation*	2.6%
	Conventional	3.7%

Q. PGE TESTIFIES THAT IT HAS BEEN LONG RECOGNIZED THAT THE THERE ARE "SERIOUS PROBLEMS" WITH THE USE OF THE SSE METRIC.[46] PLEASE COMMENT.

A. I find these comments rather puzzling. First, PGE's testimony acknowledges that SSE is quite popular with statisticians in relation to business decisions.[47] In fact, SSE has been the cornerstone of forecasting methodologies for decades. It is the very basis of regression analysis used in econometric models, such as those applied by utility companies for load forecasts, including PGE's. Exhibit ICNU/403, Falkenberg/1-3 presents pages from the workpapers for PGE's load forecast, which show reliance on SSE, not RGRMSE. Nearly all of the "forecast infrastructure" (software, educational programs, etc.) available for analysis of such models relies on SSE or similar metrics as their basis.

Second, PGE has *never* relied upon the RGRMSE metric until now. PGE has never presented any evidence to the Commission in any proceeding that relied

[46] PGE/300, Tinker-Weitzel/3.
[47] Id. at 8.

upon the RGRMSE methodology (or the permutation test for that matter).[48/] Given the potential applicability of such methods to load forecasting models used in the IRP, this strikes me as rather questionable. At the very best, PGE's use of the RGRMSE metric is "ad-hoc."

Finally, PGE has been selective in its discussion of academic articles. While PGE may give the impression that academics view SSE as flawed and RGRMSE as the "best method," that's not the case. A commentary discussing the Fides[49/] paper states "*For a single series it is perfectly reasonable to fit a model by least squares (as statisticians customarily do) and evaluate forecasts from different models and/or methods by the mean square error (MSE) of the forecasts.*"[50/] Although that commentary was critical of the use of SSE for comparisons of multiple series, the issue we are dealing with in this case is a single series. This relates to the problem of scale, which is simply not applicable to our issue.

Q. PLEASE EXPLAIN.

A. If one were trying to compare results of two different forecast models for two different series, scale would matter using SSE. Quoting results from one series in pounds and another in dollars shouldn't make a difference in evaluating methodologies, but SSE could show a difference for this reason.[51/] Since we have only one series to compare, the problem of scale is irrelevant.

48/ ICNU/402, Falkenberg/8 (PGE Response to ICNU DR 5.32). Note that PGE objected to answering this question.

49/ Robert Fildes, The Evaluation of Extrapolative Forecasting Methods, 8 International Journal of Forecasting, 81-98 (1992).

50/ Chris Chatfield, A Commentary on Error Measures, 8 International Journal of Forecasting, 99-111 (1992).

51/ Id.

Q. PLEASE EXPLAIN WHY SSE IS THE PROPER METRIC FOR THIS CASE.

A. It should not be forgotten that the ultimate goal is to predict Net Power Costs ("NPC"). In NPC calculations, errors in the FOR projections induce an additive, not multiplicative, effect. For example, the error introduced by "missing" the outage rate for Colstrip 3, is MW_{Col3}*8760*E*R, where MW_{Col3} is the Colstrip 3 capacity, E is the error in the FOR, and R, is the cost of replacement energy less the cost of Colstrip's energy. The <u>sum</u> (not the product) of all of the unit forecast errors produces the total error due to the outage rate forecast. As a result, it should be clear that the greatest impact on cost is due to the units having FORs that are the most unpredictable. Examination of the PacifiCorp data shows that some units have far more variability in the FOR series than others.[52/] These "hard to predict" units are the most responsible for NPC forecast errors. If we seek to obtain the most accurate NPC forecasts, it makes more sense to rely on the SSE metric, because it does give more weight to the inputs that cause the greatest amount of forecast error. The RGRMSE metric does not do so, and isn't useful. SSE also has a useful interpretation for business decisions and can be converted into a cost measure.

Q. IS THIS CONCLUSION CONFIRMED IN THE LITERATURE CITED BY PGE?

[52/] One of the units showed a difference between the 90th and 10th percentiles of only 3.4%, while another showed a range of 8.4%.

A. Yes. While the SSE metric provides for a clear interpretation for business decision making (providing insight in the cost of mistakes) the RGRMSE is viewed as poor in this regard because it deals with abstract percentage changes in errors.[53]

Q. PGE ALSO SUGGESTS THAT THE SSE METHOD IS "LESS RELIABLE" THAN THE RGRMSE METHOD. PLEASE COMMENT.

A. Reliability deals with repeatability of results. In ICNU Data Request 5.27, I asked PGE if it had any specific concerns related to this issue. PGE referred back to the response to ICNU Data Request 5.24, which simply stated PGE was concerned about this issue. ICNU/402, Falkenberg/9-12. In this instance, PGE proffers an un-testable and unverifiable hypothesis. Given the lack of specifics, I think this argument is esoteric and academic at the very best.

Other Issues

Q. NONE OF YOUR ANALYSES DEAL WITH THE OPUC COLLAR OR OTHER COLLARS THAT DEPEND ON THE NERC DATA. PLEASE EXPLAIN WHY.

A. Despite having close to two years to obtain the data from NERC, none of the parties has obtained data appropriate for all applicable plant sizes. Data is not available for any plants less than 300 MW in size. These are 9 of the 26 (35%) PacifiCorp units. Further, NERC data going all the way back to 1989 is not available either. Instead, PGE has data only to 1999. Consequently, the serious lack of data precludes a meaningful analysis of the forecasting accuracy gains possible using a NERC collar as compared to the ICNU collar. None of the tests reported on Table 2 of PGE/300 (which used the NERC data) are correct due to the flaws in the PGE permutation tests, or the use of the incorrect "data to date"

---

[53] Armstrong and Collopy, supra, at 74, 77.

analysis as well as reliance on the RGRMSE. Finally, it does not appear that PGE has done an analysis of the OPUC collar, which combines the 90/10 collar range (based on NERC data) with the long term average (based on own plant data) replacement strategy.

Even accepting PGE's results as accurate, Table 2 in PGE/300 merely indicates that one can't decide among the alternatives, not that one is better than the other. Consequently, the OPUC would have to accept a NERC data based collar on grounds other than statistical analysis.

Q. DO YOU HAVE ANY RESERVATIONS ABOUT THE MANNER IN WHICH THE NERC COLLAR (AS PROPOSED BY PGE) IS CONSTRUCTED?

A. Yes. The PGE construction of the NERC collar relies on the four-year average of the 90th and 10th percentiles to determine outliers. This approach is questionable because it implies that to be excluded from the four-year average; a generator would have to have an outage rate that on average placed it in the very best or very worst generators in the NERC database for four-year in a row. Typically, generators will not consistently fall outside of the industry norms for an extended period of time. While unit A, may be in the worst 10 percent in year one, in years two through four it may be somewhere in the middle. Consequently, a problem with the current construction is that it would likely produce a collar range that is "too loose." A more reasonable approach would be to obtain the four-year average FOR for each unit in the NERC sample and then determine the 90th and 10th percentiles for the four-year averages for all of the units.

Q. PGE CITES AN ARTICLE[54] WHICH IT SUGGESTS SUPPORTS REPLACEMENT OF THE OUTLIER EVENTS WITH "BOUNDARY VALUE" DATA. PLEASE COMMENT.

A. This article assumes the series in question is non-stationary. It relies on an "exponential smoothing" methodology which no party to this case has ever even suggested is applicable to forecasting outage rates. This is a fundamentally different methodology than a four-year moving average, advocated by PGE since the very start of this case. As shown above, a four-year moving average model assumes a stationary data series. I fail to see any relevance of this article to the issue at hand.

Idaho Power Testimony

Q. DOES IPC SUPPORT THE ICNU COLLAR MECHANISM?

A. No. While IPC states that it does routinely remove "extreme events" from FOR calculations, it apparently follows an ad-hoc, judgmental process. IPC believes that because it has few units and operates in two states, the ICNU collar is not useful. Further, IPC contends that data from 20 years ago is not as useful as more recent data in preparing forecasts.

Q. PLEASE COMMENT.

A. IPC should use whatever method is approved by the OPUC for PGE and PacifiCorp. Use of a judgmental collar mechanism will simply lead to more subjectivity in the outage rate forecasting process. Also, it makes little sense to believe that the Commission should adopt a different forecast method for IPC than PacifiCorp. IPC owns a share of the Bridger plant. It makes little sense to think

---

54/ Sarah Gelper, Roland Fried and Christophe Croux, Robust Forecasting with Exponential and Holt-Winters Smoothing (Department of Decision Sciences and Information Management, Katholieke, Universiteit Leuven 2007).

the two companies should have different forecasts for the same unit. Further, PacifiCorp operates in five states. This fact has no bearing on which forecasting methodology is best. Finally, IPC's objections related to use of older data is just another variation on the argument that the FOR series is non-stationary, which I have already dealt with.

Q: ARE YOU ADDRESSING OTHER ISSUES IN THIS PROCEEDING?

A: No. The issues for my reply testimony are limited to responding to PGE and IPC. In addition, I understand that PGE continues to support the settlement resolution of the non-collar related issues. ICNU/402, Falkenberg/13-14.

Q. DOES THIS CONCLUDE YOUR TESTIMONY?

A. Yes.

BEFORE THE PUBLIC UTILITY COMMISSION

OF OREGON

UM 1355

In the Matter of	)
	)
THE PUBLIC UTILITY COMMISSION	)
OF OREGON	)
	)
Investigation into Forecasting Forced	)
Outage Rates for Electric Generating Units.	)

ICNU/401

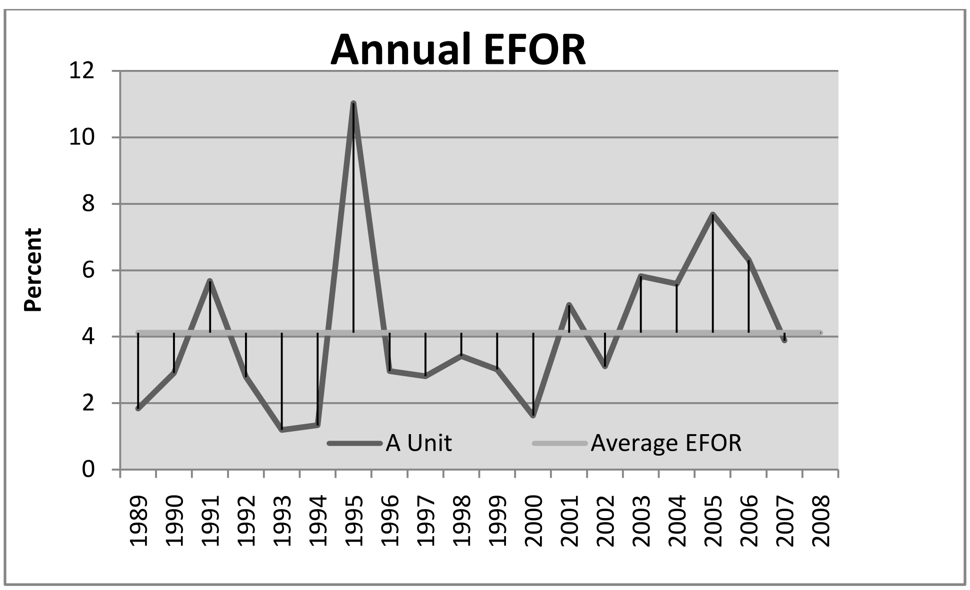
TIME SERIES ANALYSIS FORCED OUTAGE RATE DATA

August 13, 2010

TIME SERIES ANALYSIS OF PACIFICORP FORCED OUTAGE RATE DATA

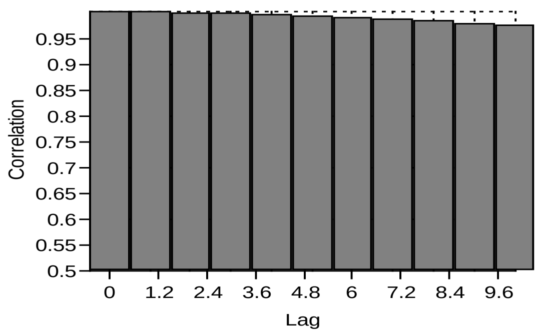
PURPOSE: This analysis tests whether the FOR series data is stationary or non-stationary. If stationary, then the assumption that the series is mean reverting is valid. Also, it would imply that because the mean does not change over time, the use of ex-ante data did not bias the results of the forecast accuracy comparisons.

METHODOLOGY: There are a number of methods available to determine is a time series is stationary. The simplest and most obvious is to simply look at a chart of the data over plotted against time. The chart below shows the data for one of the resources included in the PacifiCorp dataset. The chart shows no obvious trends in the data, suggesting the series is stationary. Typically, the results do not appear as clear-cut as this one, however.

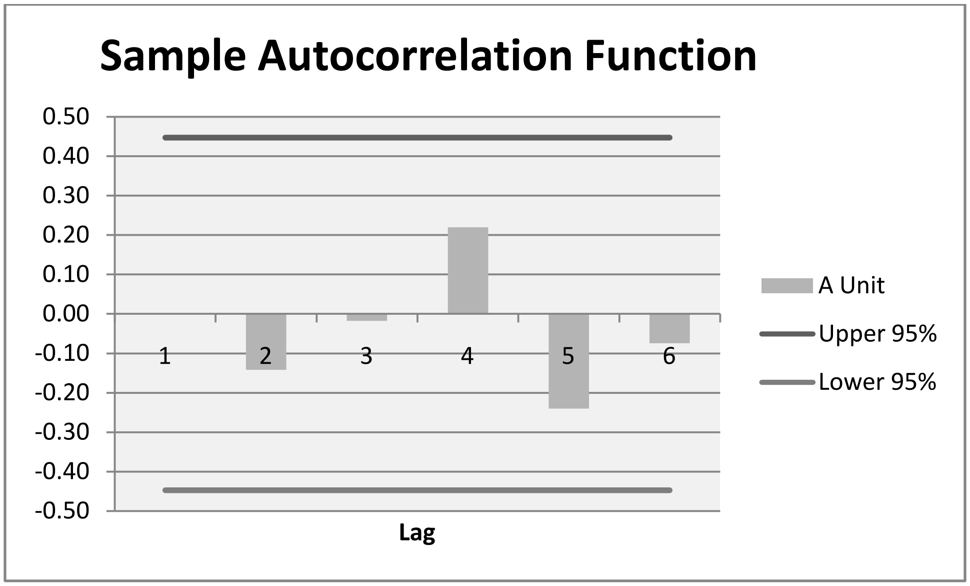


In some cases, trends may appear to exist, so more rigorous methods are required. A standard technique is to measure the correlation of the time series data with itself at various time lags. The autocorrelation of a series at lag k, is a measure of the correlation between observations Y(t) with observations Y(t-k). The autocorrelation chart presents the results for a given series at various lags. For a non-stationary series (such as most economic data) observations in one period are highly correlated to observations in prior periods. The chart below, showing the autocorrelations for the U.S. CPI from 1957 to 2009 illustrates the pattern characteristic of highly correlated, non-stationary data. CPI data is highly correlated with prior observations because each year's CPI is the product of all prior year's CPI observations.

Autocorrelations Graph: U.S. CPI Data (1957-2009)



For series that moves randomly through time, there is little correlation between present and prior observations. In this case below there is no indication of correlation between the data series at various lags. None of the autocorrelations lie outside of the 95% confidence interval and in fact, none are even close. Clearly, there is no evidence in this instance to support the assumption that the outage rate series is not stationary.



Data Requirements and Analysis

Time series analysis requires a substantial amount of data. The charts above show results for a twenty year period. However, this provides only 20 annual data points, which may not be enough to perform a valid analysis. To address this problem, much larger samples were analyzed. The monthly FOR data provided by PacifiCorp spanning the period 2003-2007 provided a sample of 60 data points. This data has previously been presented in this case as part of the analysis in ICNU/100 at pages 48-49. A second data set, based on PacifiCorp outage event data for the period 1979-2004 was also analyzed. From this dataset, 52 semi-annual outage rates were computed. Finally, a sample of 30 years of annual outage rate data for 1979-2008 was examined, again based on the PacifiCorp data.

Data was examined using GNU Regression, Econometrics and Time Series Library, "Gretl"[1/] and "PAST" a statistical package available from the Natural History Museum

---

1/ Giovanni Baiocchi and Walter Distaso, GRETL: Econometric Software for the GNU Generation, 18 Journal of Applied Econometrics 105-110 (2003). Gretl was written by Allin Cottrell (Department of Economics, Wake Forest University).

of the University of Oslo.[2/] Results were also compared to those computed using EView6, a commercial statistical package. Autocorrelations for up to six lags were computed using excel as well. These analyses verified the Gretl and PAST model produced consistent results. The Gretl model was easiest to work with, so it was used for the final results.

Following Box and Jenkins, a "portmanteau"[3/] test of the autocorrelations viewed as a whole was used. The Ljung-Box statistic[4/] was computed for the residuals of the fitted model[5/] to test for stationarity:

$$Q = n\,(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

This test compares the Q-statistic, as defined above to a "critical value" which indicates statistical significance. If the Q-value is below the critical value, the null hypothesis, that the time series is stationary cannot be rejected.

Results

Gretl was applied to the monthly, semi-annual and annual data samples. The following table shows the Q-statistic results for the annual data. None of the results approach the critical values, indicating all of the outage rate series are stationary.

The monthly data showed very little evidence of autocorrelation in most of the series analyzed. In the few where significant autocorrelation was present, the data was fit to an appropriate, single lag AR or MA model, and the residuals examined. In all cases, the residuals

---

2/ Øyvind Hammer, David A. T. Harper, and Paul D. Ryan, PAST: Paleontological Statistics Software Package for Education and Data Analysis, 4 Palaeontologia Electronica (2001).

3/ George P. Box and Gwilym M. Jenkins, Time Series Analysis: Forecasting and Control 290 (1976).

4/ G. M. Ljung and G. E. P. Box, On a Measure of a Lack of Fit in Time Series Models, 65 Biometrika 297–303 (1978).

5/ Simply the observation less the mean for series that showed no need for a more complex ARMA model, and usually and AR(1) or MA(1) model for others.

showed Q-statistics far below the critical value. This demonstrates that in the short run, recent data indicates no evidence of non-stationary behavior in the FOR data.

In about half of the cases for the longer data series there was no significant evidence of autocorrelation among the FOR series, exhibiting patterns roughly similar to that shown above. In the remaining cases where significant autocorrelation was apparent in the FOR series, the series was modeled using an AR(1) model. In all those cases, the residuals showed an absence of autocorrelation. For AR(1) models, the coefficients were in the range $-1 < \square < 1$, implying a stationary model.[6/] Consequently, the lack of autocorrelation among the residuals or in the original series supports the conclusion that the series are stationary. The workpapers provide the results of the various analyses.

[6/] George C. Tiao, An Introduction to Applied Time Series Analysis (1975); George P. Box, and Gwilym M. Jenkins. Time Series Analysis: Forecasting and Control 56 (rev. ed. 1976).

Summary of Time Series Analysis Results

	Period	*Monthly*	*Semi Annual*	*Annual*
	Start	*2003*	*1979*	*1979*
	End	*2007*	*2004*	*2008*
	N	*60*	*52*	*30*
	Qcritical	*21.0/19.7*	*18.3/16.9*	*12.6/11.1*
Unit				
1		7.1	12.7	6.1
2		12	12	4.2
3		8.1	6.9*	4.2*
4		15.3	9.7	6.5
5		7.5*	7.3	4.4*
6		16	7.4*	4.3*
7		6	4.4	5.7
8		5.5	3.3	1.3*
9		16.7	7.5	7.9*
10		17.9	11*	1.1
11		14.5	8.1*	4.7
12		4	7.9*	3.9
13		8.3	7.6*	2.7
14		7.2	7.6*	2.1*
15		8	4.1*	7.2*
16		11.7	4.1	4*
17		6.8	10.2*	0.5*
18		5.7		
19		8.6		
20		9.6*		
21		14.2		
22		12.3		
23		9.4*		
24		7.2		
25		15.1		
26		18.4		

* - Denote Q-test for simple 1 lag AR or MA model, lower Q-value applies
Note- Unit coding varies from column to column.

CONCLUSION: Based on monthly, semi-annual and annual data, the FOR series is stationary and fluctuates about its mean, or can be represented by a simple stationary AR or MA model. This analysis invalidates PGE's first two arguments[7/] listed in the summary of ICNU/400.

[7/] PGE's arguments are: 1) That it was illogical to assume the FOR series was mean reverting and; 2) That using the long term mean data produced an unfair bias in the forecast accuracy comparisons.

BEFORE THE PUBLIC UTILITY COMMISSION

OF OREGON

UM 1355

In the Matter of	)
	)
THE PUBLIC UTILITY COMMISSION	)
OF OREGON	)
	)
Investigation into Forecasting Forced	)
Outage Rates for Electric Generating Units.	)

ICNU/402

PGE AND PACIFICORP RESPONSES TO ICNU DATA REQUESTS

August 13, 2010

UM-1355/PacifiCorp
July 27, 2010
ICNU Data Request 9.1

ICNU Data Request 9.1

Please provide a list of the initial on-line date (year only is sufficient) for each of the coal-fired generating units which PacifiCorp has an ownership interest in.

Response to ICNU Data Request 9.1

The Company objects to this request on the basis that PacifiCorp has not filed testimony in this phase of the proceeding. Notwithstanding this objection, the Company responds as follows.

Coal-Fired Unit	Commercial Start Year
Carbon Unit 1	1954
Carbon Unit 2	1957
Cholla Unit 4	1981
Colstrip Unit 3	1984
Colstrip Unit 4	1986
Craig Unit 1	1980
Craig Unit 2	1979
Dave Johnston Unit 1	1959
Dave Johnston Unit 2	1960
Dave Johnston Unit 3	1964
Dave Johnston Unit 4	1972
Hayden Unit 1	1965
Hayden Unit 2	1976
Hunter Unit 1	1978
Hunter Unit 2	1980
Hunter Unit 3	1983
Huntington Unit 1	1977
Huntington Unit 2	1974
Jim Bridger Unit 1	1974
Jim Bridger Unit 2	1975
Jim Bridger Unit 3	1976
Jim Bridger Unit 4	1979
Naughton Unit 1	1963
Naughton Unit 2	1968
Naughton Unit 3	1971
Wyodak Unit 1	1978

July 23, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 4.4
Dated July 19, 2010
Question No. 007

Request:

Did PGE make any effort to obtain the full history of plant outage rates from PacifiCorp?

Response:

No. The scope of PGE's testimony was restricted to addressing ICNU's collaring and mean replacement proposal,[1] which was based on a maximum of 20 years of data for any PacifiCorp plant.

g:\ratecase\opuc\dockets\um-1355\dr-in\icnu to pge\finals\dr_007.doc

[1] Order No. 10-157, Page 5, states that "PGE may file additional testimony to address ICNU's FOR proposal."

July 23, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 4.32
Dated July 19, 2010
Question No. 035

Request:

For long term planning studies, such as the IRP, does PGE assume that the outage rates for fossil power plants will remain constant over the forecast horizon, or does PGE assume that they will be a non-stationary series and attempt to forecast the outage rates over time?

Response:

PGE objects on the basis that this subject is not relevant to this phase of the docket. Subject to and without waiving its objection, PGE responds as follows:

In PGE's IRP, outage rates for fossil power plants are long-run planning assumptions and, as such, no assumptions are made about the time series properties of the forced outage rate series.

July 23, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 4.33
Dated July 19, 2010
Question No. 036

Request:

Provide the outage rates assumed by PGE for its current IRP studies for all fossil units.

Response:

For our current resources, the percent forced outage rates are shown below:

Name	Forced Outage
Colstrip 3	10
Colstrip 4	10
Beaver CC -- 1-7	14
Boardman 1	8.5
Coyote Springs CC	3
Beaver CC -- 8	14
Port Westward Power Plant	5

The new resource percent forced outage rates are shown below:

Name	Forced Outage
Biomass	10
CCCT	4
Distributed Generation	0.1
Geothermal	10
IGCC in WY	15
Nuclear (large size) in Idaho	4
Reciprocating engines	0.3
SCP Coal in WY	8
Simple Cycle (LMS 100)	2
CHP	0.3

July 26, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 5.34
Dated July 19, 2010
Question No. 072

Request:

Assume a Moving Average (MA) model of order q, as specified below. Explain why this would not be a mean reverting series.

> $X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$ where μ is the mean of the series, the θ1, ..., θq are the parameters of the model and the εt, εt−1,... are white noise error terms.

Response:

PGE objects to this request because it is not related to PGE/300 and is outside the scope of this proceeding. Without waiving objection, PGE responds as follows:

This would not be a mean reverting series if the time series itself were not stationary:

> The foundation of time series analysis is stationarity. A time series $\{r_t\}$ is said to be *strictly stationary* if the joint distribution of $(r_{t1},\ldots,r_{tk})$ is identical to that of $(r_{t1+t},\ldots,r_{tk+t})$ for all t, where k is an arbitrary positive integer and $(t_1,\ldots,t_k)$ is a collection of k positive integers. In other words, strict stationarity requires that the joint distribution of $(r_{t1},\ldots,r_{tk})$ is invariant under time shift. This is a very strong condition that is hard to verify empirically. A weaker version of stationarity is often assumed. A time series $\{r_t\}$ is *weakly stationary* if both the mean of r_t and the covariance between r_t and $r_{t-\ell}$ are time-invariant, where ℓ is an arbitrary integer. More specifically, $\{r_t\}$ is weakly stationary if (a) $E(r_t) = \mu$, which is a constant, and (b) $\mathrm{Cov}(r_t, r_{t-\ell}) = \gamma_\ell$, which only depends on ℓ. In

> practice, suppose that we have observed T data points $\{r_t \mid t = 1,\ldots,T\}$. The weak stationarity implies that the time plot of the data would show that the T values fluctuate with constant variation around a fixed level. In applications, weak stationarity enables one to make inferences concerning future observations (e.g., prediction)[1].

As stated by Box and Jenkins[2]: (Notation in original altered to match the notation in this Data Request):

> "The sequence $\theta 1, \ldots, \theta q$ formed by the weights may, theoretically, be finite or infinite. If this sequence is finite, or infinite and convergent, the filter is said to be *stable* and the process X_t to be stationary. The parameter μ is then the mean about which the process varies. Otherwise, X_t is non-stationary and μ has no specific meaning except as a reference point for the level of the process."

Using an historical mean as a replacement value for outliers, raises an important question for PGE - whether this historical mean is the "mean about which the process varies" or "has no specific meaning except as a reference point for the level of the process." If the latter, it is hard to justify its inclusion in a forecast.

While we believe this data request is useful in clarifying the issue, we believe it also makes clearer why PGE has an issue with using the mean. If the above quoted property were to be used as the basis of a (yet unspecified) test for stationarity, it would appear to require that we employ "Box Jenkins" (BJ) methods to identify the process. For PGE's Boardman plant, this would not be possible . Boardman began operations in 1980, which means that we would have a limited number of observations. More specifically, we have far too few observations to identify a process for Boardman (assuming a BJ type model is even reasonable). Similar issues arise with the Colstrip plants that came on-line in 1983 (Unit 3) and 1985 (Unit 4).

Aside from the question of stationarity, there is another issue related to the short history for Boardman. Even if the process were stationary, the short time series could result in a poor estimate of μ simply due to sampling variability. This same problem is relevant to analyses of data for Colstrip Units 3 and 4, as those plants also have short histories. This issue as far as we know was not addressed by the Commission. The variation in the sample mean would be incorporated in the forecast error.

[1] Ruey S Tsay, *Analysis of Financial Time Series*, p. 25.
[2] George Box and Gwilym Jenkins, *Time Series Analysis: Forecasting and Control*, Revised Edition, p. 9.

July 26, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 5.32
Dated July 19, 2010
Question No. 070

Request:

Identify any previous dockets where PGE has presented evidence to the Commission related to the RGRMSE and permutation test methodologies.

Response:

PGE objects to this request because it is overly broad. Without waiving objection, PGE responds as follows:

We are unaware of any previous dockets where PGE has presented evidence to the Commission related to the RGRMSE and permutation test methodologies.

July 26, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 5.27
Dated July 19, 2010
Question No. 065

Request:

Pages 8-9. Identify the specific problems we may still face with respect to scale and reliability.

Response:

See PGE's responses to ICNU Data Request Nos. 061 and 062.

July 26, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 5.23
Dated July 19, 2010
Question No. 061

Request:

Page 8-9. Explain how scale dependency would apply in the case of outage rates. Explain why different plants should a-priori be assumed to have different size forecasting errors.

Response:

Forecasting errors result from (at least) two sources: (1) bias inherent in the forecasting method and (2) the inherent variability in the outcome (FOR).

If the forecast methodology results in biased forecasts (as we expect is true for the candidates under consideration), equal biases across plants would be a very unlikely outcome. Variations in the degree of bias would therefore result in plants experiencing different sized forecasting errors.

There are a number of reasons why variability in forced outage rates should differ across plants:

- Differences in the ages of plants
- Differences in the underlying technologies
- Differences in the quality of coal used
- Differences in the types and extent of pollution control technology
- Operational differences: For example, some plants might provide some level of ancillary services while others do not

- Differing maintenance philosophies (tradeoff between costs of more frequent maintenance and costs associated with outages).

As we discuss more fully in our response to ICNU Data Request No. 063, we want to be able to make valid inferences for other populations (PGE plants) from Mr. Falkenberg's results:

> "The scale of the data often varies considerably among series. Series with large numbers might dominate comparisons"[1]

g:\ratecase\opuc\dockets\um-1355\dr-in\icnu to pge\finals\dr_061.doc

[1] J. Scott Armstrong and Fred Collopy, "Error Measures for Generalizing About Forecasting Methods: Empirical Comparisons", *International Journal of Forecasting*, 8 (1992), 70.

July 26, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 5.24
Dated July 19, 2010
Question No. 062

Request:

Page 8-9. Explain how the referenced "reliability" considerations factor into this analysis of outage rate forecasting.

Response:

"Reliability addresses the question of whether repeated application of a procedure will produce similar results"[1] Since we are very concerned with the degree to which we can generalize the results from the PacifiCorp data, reliability is a key issue in choosing a method to aggregate forecast errors across plants.

[1] J. Scott Armstrong and Fred Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons", *International Journal of Forecasting* 08 (1992) 69-80.

August 2, 2010

TO: Irion A. Sanger
Industrial Customers of NW Utilities

FROM: Randy Dahlgren
Director, Regulatory Policy & Affairs

PORTLAND GENERAL ELECTRIC
UM 1355
PGE Response to ICNU Data Request 6.1
Dated July 27, 2010
Question No. 073

Request:

Does PGE still support the Staff/PGE/ICNU/CUB settlement's proposed resolution of the issues? Specifically does PGE continue to support the settlement's proposed resolution of:

1) **Applying the EFORd concept to Beaver Units 1-7, and Beaver Unit 8;**
2) **The Split Maintenance Outage Rate between on- and offpeak hours for Colstrip and Boardman;**
3) **Adoption of Staff's proposed formulas for rates of forced outages, planned outages, and maintenance outages, or a showing of equivalence to formulas used; and**
4) **Adoption of Wind reporting requirements.**

Finally, does PGE agree that the issue of Planned Maintenance Outage should not be resolved in UM 1355, but in a separate docket(s).

If PGE no longer supports the settlement proposed resolution of these issues, then please explain.

Response:

PGE objects to this request because the subject is not part of this phase of the docket and is not discussed in PGE/300. We note that this phase of the proceeding is limited to discussion of the proposal made in ICNU/300. Specifically, Order No. 10-157 states that, in this phase, "PGE may file additional testimony to address ICNU's FOR collar proposal." (Order No. 10-157, Page 5). Without waiving its objection, PGE responds as follows:

PGE agrees that planned maintenance outage (PMO) issues should not be resolved in UM 1355, as they were already resolved in UE 208. In UE 215, PGE continues to adhere to the UE 208 resolution of PMO issues.

PGE supports items 1 through 4 of the request as a package. We are following all four of these proposed resolutions in the net variable power cost portion of UE 215.

The REG Procedure
Model: MODEL8
Dependent Variable: NMF

Number of Observations Read	322
Number of Observations Used	250
Number of Observations with Missing Values	72

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.855125E12	6.855125E12	2.671E8	<.0001
Error	249	6389795	25662		
Uncorrected Total	250	6.855131E12			

Root MSE	160.19310	R-Square	1.0000
Dependent Mean	161762	Adj R-Sq	1.0000
Coeff Var	0.09903		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
NMF_LG		1	0.99996	0.00006118	16344.2	<.0001

FORECAST(COMCUST)
GROWTH IN SCHEDULE 32 CUSTOMER

The REG Procedure
Model: MODEL1
Dependent Variable: G_N32

Number of Observations Read	310
Number of Observations Used	309
Number of Observations with Missing Values	1

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5831113	5831113	106.99	<.0001
Error	308	16785892	54500		
Uncorrected Total	309	22617005			

Root MSE	233.45160	R-Square	0.2578
Dependent Mean	138.78641	Adj R-Sq	0.2554
Coeff Var	168.20927		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
G_NSC7	1	0.11612	0.01123	10.34	<.0001

```
UPVMHNH
Restricted Cochrane-Orcutt
MONTHLY data for  295 periods from  JAN 1985 to  JUL 2009
Date: 16 SEP 2009

upvmhnh

     - 497.428 * temp_sq*winter - 453.830 * temp_sq*spring
      (32.4605)                  (25.9072)

     - 22.0155 * temp*swing - 24.6202 * temp*summer
      (13.6385)               (4.83622)

     + 12599.8 * winter + 12271.9 * spring + 10175.8 * swing
      (5.47333)          (5.33495)           (4.43671)

     + 57.6346 * jan + 1.36908 * c1d65
      (5.74638)        (6.32830)

     + 11.5541 * wind*(winter+spring) - 1.32253 * unemp_or[-1]
      (6.31015)                         (2.26906)

     - 1.76338 * unemp_or[-2]
      (2.26906)
                                 - 1.32253 * unemp_or[-3]
                                  (2.26906)

     + 5.17979 * trend*summer + 38.4478 * step(2003,5)
      (4.51111)                (3.88147)

     - 0.91197 * p2tariff/cpi - 1.36796 * p2tariff/cpi[-1]
      (0.47781)                (0.47781)

     - 1.36796 * p2tariff/cpi[-2] - 0.91197 * p2tariff/cpi[-3]
      (0.47781)                    (0.47781)

     - 8127.56
      (3.53732)

   ynomial lags:

    unemp_or
    from  1 to  3 degree  2 near far

    p2tariff/cpi
    from  0 to  3 degree  2 near far


Sum Sq      224161   Std Err    28.3960   LHS Mean  945.028
R Sq        0.9903   R Bar Sq    0.9897   F 16,278  1769.86
D.W.( 1)    1.9115   D.W.(12)    1.8599

AR_0 =   + 0.25959 * AR_1 + 0.43855 * AR12
          (4.51377)        (8.13630)
```