GLOSSARY

TERM	DEFINITION
Adjusted R-Squared Corrected R-Squared	Measures the proportion of the variation in the dependent variable accounted for by the estimated model. The R-squared value is sensitive to the number of independent variables included in the regression model. The addition of more independent variables to the regression equation can never lower the R-squared value and will reduce the degrees of freedom of the model. An Adjusted R-squared value is a better measure of "goodness of fit" of a regression model than the R-squared value as it compensates for the number of independent variables included in the regression model.
Autoregressive Integrated Moving Average ("ARIMA") (Box-Jenkins) AR (p) Terms MA (q) Terms	A short-term forecasting technique based on identifying patterns in the history of the time series and extending those patterns into the future. Unlike a regression model, ARIMA models do not use explanatory variable to explain changes in a time series. Instead, changes in a time series are linearly related to their own past values and to a weighted sum of current and lagged random disturbances. The model specification (specifying the number of autoregressive and moving average terms in a model) of a time series is applicable for a stationary series. The model specification (specifying the number of autoregressive 'p' lags ("AR (p)") and moving average 'q' lags ("MA (q)") terms in a model) of a time series is applicable for a stationary series. If a series is not stationary, it can be converted to a stationary form by differencing it one or more times. These models are generally useful for short-term forecasting because, for longer lead periods, past observations have little or no effect on the forecast and the forecasts tend to approach the constant or mean value of the series.
Autocorrelation	A measure of interdependence of observations at certain lag periods (e.g., "k" periods apart) It is the correlation between the original values of the series and its k-period-apart values. An autocorrelation value is considered significant if it is greater than its two-standard-deviation value. While the usage of the term autocorrelation is more common in time-series models and can be applied to any series, the use of the term "serial correlation" is more common in multiple-regression models and is applied to residual series only (which is a series resulting from the difference between the actual values and the estimated values from the model). The plot of the autocorrelation values for values of k = 1,2,3, etc. is called autocorrelation function ("ACF").
BBtu/MMBtu	BBtu equals one billion British thermal units ("Btu"). MMBtu equals one million Btu.

Billed Sales	Billed sales in a month pertain to gas consumption based on the customers billed in that month. A local distribution company ("LDC") reads customer meters on various billing cycles in a month and bills its customers within one or two business days after the meters are read. Therefore, all the meters read and billed in a month do not measure the gas used by the customer in that month. Billed sales in a month include certain volumes of gas used in the prior month and certain volumes of gas used in the current month.
Calibration Ratio	The proportion of a forecast of a base year from one modeling approach (e.g., annual forecasts) to the forecasts of the same year from another modeling approach (e.g., daily forecasts). Calibration ratios are applied to one set of forecasts (daily forecasts) to obtain consistent forecasts with the other set of forecasts (annual forecasts).
Causal Model	Used to identify a statistical relationship between a dependent variable and a set of independent or causal variables. In causal models, changes in a single variable are related to a set of logical and economic causal variables. For example, consumption of food in a family is driven by the number of persons constituting the family and their income. Causal models include econometric models as well as multivariate time-series models.
Chow Test	A statistical test devised by Gregory C. Chow to test the stability of estimated parameters of a regression model over the entire data range used in estimating the parameters. The Chow test is conducted by splitting the original data range in half, estimating the same equation on each subset, and determining if the coefficients from the two equations are statistically equal.
Cold Snap	A prolonged series of days at or near design conditions. The purpose of a cold snap is to test an LDC's ability to meet demand under a prolonged period of cold weather.
Company Use	Gas volumes consumed by the Company for its internal use.
CUSUM AND CUSUMSQ	Plots of cumulative residuals and cumulative residual squares, respectively, often used to detect changes in a time series.
Degree Day or Heating Degree Day ("HDD")	A degree day indicates how far a day's average temperature departs from 65 degrees Fahrenheit without regard to wind speed in determining the coldness of the weather. For example, a day when the mean temperature is 35 degrees Fahrenheit would be 30 HDD.
Degrees of Freedom	The number of observations in a data set used to estimate a model minus the number of constraints placed on the data set. In the context of multiple regression, degrees of freedom are equal to the number of observations used in estimating a model minus the number of parameters estimated. Degrees of freedom higher than the number of parameters estimated is recommended as it provides a more unconstrained data set to evaluate the goodness of the estimated model.

Dependent Variable	The variable to be forecasted or modeled. The variables used to explain the changes in the dependent variable are called independent variables or explanatory variables.
Design-Year Scenario	Accounts for extreme weather conditions. This scenario is developed to keep the probability of exceeding the design conditions very low (around five percent) while still striking a balance between the cost of providing supplies at higher standards and potential damage to customers in case of not receiving gas supplies.
Differenced Form	A method of transforming data by taking the differences of the original data.
Dummy Variable	A binary variable whose value equals one when a specific time-related event occurs and equals zero otherwise. These variables are created and used in modeling to represent time-related events that are difficult to quantify. • A point dummy variable represents a single abnormal data point in a series. • A seasonal dummy variable measures the effect of a particular season. It is created when a cluster of periods in each year is assigned the same binary values. • A shift dummy variable measures the effect of qualitative events, like the reclassification of customers, application of new rules and regulations, etc. • An interaction dummy variable measures the joint effect of a dummy variable with another non-dummy variable, like a seasonal dummy and number of customers.
Durbin - Watson ("DW") Statistic	A statistic used to test the significance of serial correlation at lag 1 in the residual series from a regression model.
Effective Degree Day ("EDD")	An EDD takes into account wind speed along with HDD in determining the coldness of the weather.
Elasticity	Used to interpret the effect of a percentage change in an independent variable (e.g., income) on the percentage change in a dependent variable (e.g., consumption). Elasticity values are unbound and may be positive or negative.
Ex-post Forecasts	Forecasted values from a model based on a subset of the historical data. As a result, for the forecast period, the values of the dependent variable and all the independent variables are actual values and are thus known with certainty. This analysis is useful in evaluating a forecasting model and the forecasts produced from it.

Exponential Smoothing	A forecasting method that assumes the data consist of irregular fluctuations around a constant or slowly changing level. It involves the use of an exponentially weighted moving average model for smoothing, represented as: $S(t) = \alpha*Y(t) + \alpha*(1- \alpha)*Y(t-1) + \alpha*(1- \alpha)*(1- \alpha)*Y(t-2) +$ -------- or $S(t) = \alpha*Y(t) + (1-\alpha)*S(t-1)$ where Y(t) denotes the actual value of the series at time 't' and S(t) denotes the smoothed value of the Y at time 't' and α is a smoothing constant.
F-statistic	A statistic used to test the significance of a multiple regression by testing the hypothesis that none of the explanatory variables help to explain the variation in the dependent variable.
Firm Sendout	Gas volumes dispatched to firm-sales customers who buy gas from the Company. Includes Company use and unaccounted-for gas.
Firm Throughput	Gas volumes dispatched to firm sales and firm transportation customers including Company use and unaccounted-for gas. It is equal to throughput volumes minus the volumes dispatched to interruptible sales and interruptible transportation customers.
Heating Degree Day ("HDD")	See Degree Day.
Heteroskedasticity	A situation in a multiple-regression model where the errors do not have constant variance across the entire range of values. Its presence causes the standard error estimates of each of the estimated parameters to be biased and the statistical tests to be incorrect, and thus violates one of the basic assumptions of multiple linear regression models.
Holt Exponential Smoothing	A forecasting technique that fits a linear trend model where both the level and trend are based on smoothed estimates. Such models are useful to describe random walk processes or situations where the series has many ill-behaved transient periods.
Line or Unaccounted-for Loss	The amounts of gas that leak out through the Company's distribution system and are not accounted for by the LDC. The amount of gas lost depends on the age and type of underground pipe and the length of the Company's distribution system.
Local Distribution Company ("LDC")	A company that delivers natural gas to consumers within a specific geographic area.
Multicollinearity	A situation in a multiple-regression model where two or more explanatory variables (or a combination thereof) are highly correlated with each other. Theoretically, if any linear combination of one set of explanatory variables is nearly perfectly correlated to a linear combination of any other set of explanatory variables, then a multicollinearity problem is present. Multicollinearity causes the standard errors of the estimated parameters to be large and hence their t-values to be small, which in turn may result in erroneous conclusions in accepting or rejecting a multiple-regression model.

Multiple Regression	A statistical modeling technique where the values of a variable of interest such as sales are explained in terms of many causal or explanatory variables like population, price, disposable income, etc.
Normal Year	A standard based on the weather an LDC has experienced over a period of approximately 30 years. It represents the arithmetic average of temperatures experienced during this period.
Ordinary Least Squares ("OLS")	A procedure used to estimate the parameters of a multiple- regression equation that minimizes the residual sum of the squares.
Outlier	A rare or unusual observation whose value is far from the general range of other data points in a series. The parameter estimates of a regression model are very sensitive to the presence of outliers. If outliers are not accounted for before estimation, their presence will be seen in the residual series.
Partial Autocorrelation Function ("PACF")	Refers to coefficient values from various regressions in which a stationary series is regressed sequentially with its past values at lag 1, lag 2, lag 3, etc. until the term associated with the last lag is no longer significant. The partial autocorrelation at lag 'k' is defined as the coefficient associated with the k period lag term when Y is regressed on Y(t-1), Y(t-2), …., Y(t-k) (i.e., the k'th coefficient of the k'th regression). The plot of the partial autocorrelation values for values of "k" = 1, 2, 3,… is called partial autocorrelation function ("PACF").
Partial Correlation	Measures the correlation between two sets of values of a series which are "k" periods apart when the effect of all intermediate lags are removed. In other words, it measures dependency between these two sets of observations which is not accounted for in other earlier lags.
R-Squared	Also called the "goodness of fit" value. Measures the proportion of the total variation in the dependent variable being explained by the set of causal or explanatory variables. It varies between zero and one, and a value closer to one with all other characteristics of the model being statistically and logically sound indicates a good model.
Residual Series	A series representing the difference between the actual value of the variable being modeled and the estimated value based on the model. Generally, a good regression model is one that tries to explain or account for a large portion of the variance in the dependent variable. If the estimated model has captured or explained all the information content in a data set, then the residual series should behave like a random series showing no auto or partial correlations at various lags. Presence of large residuals implies a poor fit, while presence of small residuals implies a good fit.
Send Out	The total volume of gas distributed by a company, including gas used by the company and gas unaccounted for.

Serial Correlation (Autocorrelation)	A situation where the error terms in a multiple-regression model are correlated at different time periods. It is similar to correlation between two variables but relates to the series at different time lags. Its presence causes the standard error of the parameters to be biased downwards, thus causing the t statistics to be greater than they should be and, therefore, may result in erroneous decisions regarding the variables to be accepted in the model.
Significance Level	The criterion used for rejecting a null hypothesis in hypothesis testing. In regression models, the significance level is used in deciding whether an independent variable included in the regression equations should be retained in the model. First, a coefficient associated with the independent variable is estimated and the difference between the estimated coefficient value and the null hypothesis value is determined. The null hypothesis in a regression model assumes that the coefficient value is zero. Then, assuming the null hypothesis is true, the probability of the difference (equal to the estimated value of the coefficient) is computed. Finally, this probability is compared to the significance level. If the probability is less than or equal to the significance level, then the null hypothesis is rejected, the independent variable is retained in the model, and the outcome is said to be statistically significant. Traditionally, econometricians have used either the 0.05 (five percent) level or the 0.01 (one percent) level, although the choice is largely subjective. The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 0.01 level is more conservative than the 0.05 level. The Greek letter alpha (α) is sometimes used to indicate the significance level.
Standard Error of Coefficient	A measure of the dispersion of the estimated coefficient value about its mean.
Standard Error of Regression	A measure of the standard deviation of the error term in the regression model.
Stationary Series	Signifies a series that has the same statistical behavior at each point in time and moves randomly about its mean and within a fixed band. Statistically, the probability density function of a stationary series does not change with time and, therefore, its mean and variance also do not change with time.
t-statistic	A statistical test used in multiple-regression analysis to test the hypothesis that the individual estimated regression parameters are significantly different from zero. It is computed as the ratio of the coefficient to the standard error of the coefficient. To test the hypothesis, generally a five- or ten-percent level of significance is used. If the calculated t-value of a parameter is greater than its critical (tabulated) t-value at a five- or ten-percent level of significance, the hypothesis that the parameter value is zero is rejected.

Throughput	Refers to gas volumes that are dispatched during a time period. This includes gas used by all customers including firm and interruptible sales service customers who buy gas from the LDC, firm and interruptible transportation customers who buy gas from third party suppliers, Company-use gas, and gas lost during distribution.
Time-Series Models (univariate)	Prediction models based on extrapolating the historical patterns of a variable's history. A variable's historical data might indicate the presence of time trend, seasonality, or lagged correlations. Thus, the forecasts of a variable from time-series models are based solely on the past behavior of that variable alone and not related to changes in other influencing time series or causal variables. Such models include exponential smoothing, winter's smoothing, trend models, and ARIMA or Box-Jenkins models.
Unbilled Sales	Unbilled sales pertain to gas volumes that are used in a month or period but not billed in that month or period.
White Test	A statistical test for heteroskedasticity. It follows a Chi-square distribution and is calculated by multiplying the number of observations with the R-squared value of a regression when squares of the residual series of an original equation are regressed against independent variables in the original equation, a cross product between independent variables and squared independent variables.