Accounting for Serial Autocorrelation in Decline Curve Analysis

Eugene Morgan[a,*]

[a] *The Pennsylvania State University, John and Willie Leone Family Department of Energy and Mineral Engineering, 110 Hosler Building, University Park, PA 16802-5000*

Abstract

Decline Curve Analysis is a popular tool in the oil and gas industry for forecasting well production and estimating reserves. Current decline models fail to capture all of the behavior in shale gas production histories, due to the complex flow regimes of these low-permeability reservoirs. That is, upon fitting one of these models, one often sees significant and sustained deviation of the flow rate data points from the decline trend. One way to measure this "lost signal" is to look at the autocorrelation in the residuals about the fitted decline model. Indeed, with many shale gas wells we see significant amounts of autocorrelation, especially when comparing the flow rate at one time to the next (lag one). Theoretically, this serially autocorrelated error can impact decline curve analysis in two ways: 1) inefficient estimation of decline curve parameters, and 2) lost signal in the data. Borrowing from time series statistics, there are two conventional ways of dealing with these potential problems: 1) estimate the decline curve parameters with generalized least squares or generalized nonlinear least squares, and 2) fitting an ARMA model (or variants) to the residuals and adding it to the fitted decline curve.

This paper investigates the practical implications of these two procedures by exercising them over decline curves fit to 8,527 Marcellus shale gas wells (all wells from that play with viable data for the analysis). The study explores the effect that generalized regression methods and ARMA-modeled residuals have on six different decline curves, and performance is measured in terms of sum of squared residuals (a metric for goodness-of-fit, calculated on the training data), mean absolute percent error (a standard metric for forecasting accuracy, calculated on the testing data), and prediction interval coverage rate (to examine the efficiency of probabilistic forecasts, also calculated on the testing data). These metrics are computed on the first 24, 36 and 60 months of training data at each well, with the testing data comprised of the remaining production rates.

The main finding of this study is that the prediction intervals nearly always show better coverage of the testing data with inclusion of the ARMA-modeled

residuals, regardless of the type of decline curve used or length of production data used for fitting. Similarly, the uncertainty about EUR values is better quantified when using the ARMA-modeled residuals. Including ARMA-modeled residuals also frequently improves forecasting accuracy, especially for Duong's decline curve and for other curves when more data are used for fitting, and consistently improves the goodness-of-fit. The use of generalized least squares or generalized nonlinear least squares has little benefit in fitting the decline curves, except for the Logistic Growth model, where it improves both fit and forecasting accuracy. Overall, one can expect more accurate production forecasts with better uncertainty quantification when accounting for serial autocorrelation in the production data, which is needed for the optimal development of both conventional and unconventional resources.

Keywords: prediction intervals, generalized least squares, ARMA, forecasting methods, production optimization, shale gas

1. Introduction

Decline curve analysis (DCA) serves as a popular method for determining EUR and forecasting production. By only requiring historical production rate data, DCA has an advantage over other methods, such as reservoir simulation (Erdle et al., 2016), which requires data for reservoir and fluid properties. For this reason, DCA is commonly used for reporting reserves to the U.S. Securities and Exchange Commission for assets under production and managing petroleum resources (SPE, 2018). Thus, DCA offers an accurate and low-cost (in terms of data and computational requirements) method for modeling and forecasting oil and gas production, from both conventional and unconventional reservoirs. As examples of unconventional applications, Weijermars (2014) uses Arps' Hyperbolic decline curve to forecast U.S. shale gas production under different scenarios, and Yuan et al. (2015) feature DCA prominently in their review of economic evaluation techniques for shale gas development.

The initial decline curves were designed for conventional reservoirs by Arps (1945) in two forms: 1) the Exponential model,

$$q_t = q_i \exp(-D_i t), \tag{1}$$

and 2) the Hyperbolic/Harmonic model,

$$q_t = q_i(1 + D_i b t)^{-1/b}, \tag{2}$$

*Corresponding author

where $0 < b < 1$ is hyperbolic and $b = 1$ is harmonic. Many unconventional wells show $b > 1$. Some modern advances are designed for shale gas production, such as the Power Law Loss-ratio model (Ilk et al., 2008),

$$q_t = q_i \exp(-D_\infty t - D_i t^n), \tag{3}$$

the Stretched Exponential (Valko and Lee, 2010),

$$q_t = q_i \exp(-(t/\tau)^n), \tag{4}$$

the Logistic Growth model (Clark et al., 2011),

$$q_t = Knt^{n-1}/(a + t^n)^2, \tag{5}$$

and Duong's model (Duong, 2011),

$$q_t = q_i t^{-m} e^{\frac{a}{1-m}(t^{1-m}-1)} + q_\infty. \tag{6}$$

In all of the above equations, our independent variable is time $t = 1, 2, \ldots, T$ and our dependent variable is flow rate q_t. As you can see, these models have different coefficients, although some share common ones (e.g., initial flow rate q_i). More recent studies focus on developing new decline curves that capture the complex flow regimes of unconventional, tight reservoirs (e.g., Wang et al. (2017)), but little attention has been paid to how these models are fit to the data. That is, the technological advances of these new models could be undermined by poorly-determined values of their parameters.

Furthermore, the value of forecasts from any model is difficult to determine without some measure of confidence in those predictions. Confidence intervals about the model predictions, or "prediction intervals", estimate a range that the forecasts will fall into with some arbitrary probability. For example, in oil and gas, one is often interested in estimating P90 and P10, which are the 10th- and 90th-percentiles of predicted rates, and so define an 80% prediction interval. Accurate determination of P90 and P10 is important for risk management associated with the economic appraisal of a field (Weijermars et al., 2017). Cheng et al. (2010) point out that conventional bootstrap methods tend to underestimate uncertainty in DCA, and also suffer from assuming that the production rate data are independently and identically distributed (i.e., not autocorrelated). Gong et al. (2014) and de Holanda et al. (2018) illustrate the careful calibration of prior distributions (for the decline curve parameters) that is needed to make Bayesian probabilistic DCA successful. Both studies fit prior distributions to histograms of decline curve parameter values that are calculated beforehand using deterministic methods. Without calibration to historical data, these priors are arbitrarily and subjectively determined by the user. With the resulting prediction intervals depending strongly on these priors, the user can ultimately exert

great influence (either knowingly or unknowingly) on the estimation of uncertainty. Furthermore, such calibration may not be possible at fields with limited data, and borrowing the calibration performed at a data-rich field may not be appropriate. Thus, there is a need for probabilistic DCA methods that establish accurate prediction intervals without subjective inputs or data-intensive calibration.

Despite the inherently temporal nature of production histories, there exists little published work that applies time series statistics to modeling and forecasting of well production rates. Ayeni and Pilat (1992) apply time series statistics (ARIMA modeling) to 12 oil wells. Using monthly production data, the ARIMA models out-perform Arps' Exponential and Hyperbolic models, but this comparison is based on fits to large samples of at least 50 months. Similarly, Olominu and Sulaimon (2014) find that a particular ARIMA model does better than Arps' Exponential model when tested on one cumulative oil production curve, but the performance is based on goodness-of-fit (to all available data) and not forecasting accuracy. Gupta et al. (2014) perform ARIMA modeling on 30 unconventional wells from the Barnett, Bakken, and Eagle Ford, and find that it performs similarly to the Duong (1989) decline curve model, but it is not clear what this performance is based on and how much data is used to fit the models. Additionally, a couple studies have applied time series statistics (ARIMA modeling) to national crude oil production (Ediger et al., 2006, Yusof et al. (2010)). Cheng et al. (2010) examine the autocorrelation of residuals (after fitting Arps' hyperbolic model), but only in order to determine the appropriate block size for a block bootstrapping approach to probabilistic forecasting. No discussion is given to the effect of autocorrelated error on the estimation of the decline curve parameters, and no attempt is made to model this autocorrelation structure in the residuals. Machine learning offers another source of improvement beyond decline curves. There has been some work exploring the application of nonlinear autoregressive neural networks with exogenous inputs (NARX networks) to production data (e.g., Sheremetov et al. (2014)). In another example, Frausto-Solís et al. (2015) forecast oil production with a "Simulated Annealing based on Machine learning" approach.

The previous literature above largely focuses on conventional oil wells, and when unconventional production data are used (as in Gupta et al. (2014)), appropriate, modern decline curves are not applied (e.g., Power Law Loss-ratio, Stretched Exponential, Logistic Growth, and Duong's model). Unconventional wells exhibit much more complex production behavior than conventional wells, and a purely data-driven approach, such as ARIMA modeling alone, may not be able to capture such behavior, especially with limited production data, which is inherently non-stationary and heteroscedastic. These modern decline curves are designed to accommodate the salient trends and patterns seen in unconventional

production data in their functional forms, and can project such trends into the future with limited historical data (although, such forecasts are expected to become more reliable with more historical data). These observations motivate a hybrid approach where the decline curve models the general trend of production decline and time series modeling of the residuals about this decline curve capture deviations away from the bulk trend, as expressed through autocorrelation.

Autocorrelation, $\rho(h)$, refers to the degree which an observation at time, x_t, depends on any previous observations, x_{t-h}, where h is called the "lag". Autocorrelation can be estimated from a sample by

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \tag{7}$$

where $\gamma(h)$ is the autocovariance at lag h. It is estimated by

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}), \tag{8}$$

where $\bar{x}$ is the sample mean of x. Much of the dependency of production rates on time is described by any one of the decline curves above. However, this paper shows that significant autocorrelation can exist in the residuals after fitting a decline curve. This finding is important for two main reasons: 1) estimators of decline curve parameters that do not consider this autocorrelated error are inefficient; and 2) this autocorrelation in the residuals can be modeled and added to the decline curve fit, leading to a better performing model of production decline.

To elaborate on the first reason, consider the linear regression equation (presented here in terms of linearizing Arps' Exponential model by log-transformation):

$$\ln(q_t) = \beta' t + \varepsilon_t. \tag{9}$$

An implicit assumption in fitting this equation (by least squares or maximum likelihood) is that the error at each time, ε_t, is independently and identically distributed (i.i.d.), which leads to the ordinary least squares (OLS) estimator for the coefficients β being

$$\hat{\beta} = (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\ln(\mathbf{q}). \tag{10}$$

This assumption is broken when the residuals display autocorrelation, which implies dependency between times. While this doesn't theoretically introduce any bias in the estimation of the coefficients, it does make their estimation inefficient, which means that the variance of the estimated value is not minimal (Cochrane and Orcutt, 1949). One can mitigate this inefficiency through "generalized least squares" regression, in which the correlation structure of the error is specified. The OLS estimator above then becomes

$$\hat{\beta} = (\mathbf{t}'\Sigma^{-1}\mathbf{t})^{-1}\mathbf{t}'\Sigma^{-1}\ln(\mathbf{q}). \tag{11}$$

This is the generalised least squares (GLS) estimator of the linear regression coefficients, in which the covariance matrix Σ contains the serial autocorrelation in the off-diagonal elements (the diagonal elements contain the variance of ε_t). If Σ is unknown (which it usually is), one can estimate it by exercising Eq. 8 on the residuals from the OLS regression.

This same premise holds true for non-linear regression with nonlinear least squares (NLS) (Gallant et al., 1976), which is used in fitting all decline models presented above, except Arps' Exponential model. With nonlinear regression, let $f(t;\theta)$ represent any decline curve with vector of parameters θ. One finds the optimal parameter values $\hat{\theta}$ by minimizing

$$SSR_{NLS}(\theta) = \sum_{t=1}^{T}(f(t;\theta) - q_t)^2 = \sum_{t=1}^{T} e_t^2. \tag{12}$$

The generalized nonlinear least squares estimates of $\hat{\theta}$ factor in the correlation between error terms at different times and can be found by minimizing

$$SSR_{GNLS}(\theta) = [\mathbf{q} - \mathbf{f}(\theta)]'\Sigma^{-1}[\mathbf{q} - \mathbf{f}(\theta)]. \tag{13}$$

Minimization in either case can take place using any one of a variety of iterative, optimization algorithms (Gauss-Newton, Levenberg-Marquardt, etc.).

Generalized (nonlinear) least squares only pertains to the estimation of the regression coefficients. Even after estimating the coefficients in this manner, the residuals may still contain autocorrelation. Such structure can be modeled and incorporated in the decline curve, say in an additive way. Suppose we generically define the decline curve as $f(t;\theta)$, where θ is the vector of decline curve parameters we wish to estimate, then our regression equation is

$$\ln(q_t) = \ln(f(t;\theta)) + e_t. \tag{14}$$

If we know that e_t is at least partially composed of an autocorrelated signal, we can capture that signal with an autoregressive model of order p (AR(p)):

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + w_t, \tag{15}$$

where now w_t is i.i.d. Gaussian white noise with zero mean. Alternatively, the moving average model of order q (MA(q)), treats the signal as a linear combination of white noise terms:

$$e_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}. \tag{16}$$

The mixed autoregressive moving average (ARMA) model is then:

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}, \tag{17}$$

where $\phi_p \neq 0$ and $\theta_q \neq 0$, and the time series e_t is stationary (Shumway and Stoffer, 2010).

While these theoretical points should be taken into consideration when fitting decline curves to production data, whether they afford any practical benefit remains to be tested. This is the primary goal of this paper: to perform a thorough, comparative, quantitative analysis between the traditional regression procedure for decline curves (OLS or NLS) and 1) regression via generalized (nonlinear) least squares (GLS or GNLS), 2) additive inclusion of an ARMA model with the decline curve, and 3) the combination of GLS/GNLS with ARMA-modeled residuals. In so doing, the Methodology section presents a method of fitting decline curves and estimating prediction intervals in a manner that is free of subjectivity and *a priori* calibration.

Another goal of this paper is to demonstrate the proposed improvements to DCA on a BIG dataset. Doing so gives a more convincing validation of the research than on a small dataset, which may appear to be preferentially selected to give a favorable validation, or at least lack statistical power and generality. Working with BIG data also increases the potential for observing patterns that may not be apparent in smaller datasets, such as trends in behavior across a play, and exploring the limitations of the methodology at hand. To this end, I use all viable gas production histories from the Marcellus shale, which totals 610,192 monthly rate observations from 8,527 wells.

2. Methodology

The goal of this paper is to investigate the effect that the inclusion of autoregressive and/or moving average terms may have on decline curve performance, not only in terms of decline curve parameter estimation via generalized least squares (GLS) or generalized nonlinear least squares (GNLS), but also in terms of ARMA modeling of the residuals. The overall methodology to make these comparisons is as follows:

1. Fit decline curve by OLS (Arps' exponential) or NLS (hyperbolic, logistic growth, power law loss ratio, stretched exponential, and Duong's model) to training data, q_{train}
2. Calculate residuals, $e_{train} = \ln(q_{train}) - \ln(\hat{q}_{train})$, where $\hat{q}_{train}$ are the predictions from the fitted decline curve
3. Iteratively fit ARMA models of varying order to the residuals to get optimal orders, p and q
4. Re-fit decline curve by GLS (exponential model) or GNLS (all other models) with ARMA(p,q) correlation structure
5. Repeat step 3 on residuals from GLS/GNLS fit

Thus, there are four distinct outputs whose performance needs to be evaluated: 1) the base model from step 1, 2) the base model combined with the ARMA-modeled residuals from step 3, 3) the GLS/GNLS fit model, and 4) the GLS/GNLS fit model with the ARMA-modeled residuals from step 5. To evaluate performance, all regressions in this paper are performed on the first T_{train} months of data from each well for the training data, and the fitted decline model forecasts are compared to the remainder of the record $(T_{train}+1,\ldots,T)$ at each well. In practice, one would use all available data for fitting.

To elaborate more on these steps, in the first step, Arps' exponential model can be fit with OLS after taking the natural log transform of Eq. 1. However, all other decline curve models considered in this paper cannot be linearized and required NLS in order to estimate the decline parameters. (While Duong's model is conventionally fit with a stepwise OLS procedure, where OLS is used on a reduced form of the model to get a and m and then used again once these two parameters are known to get q_i and q_∞ (Duong, 2011), I show in this paper that using NLS on Eq. 6 above gives a better fit to the data.) Nevertheless, performing nonlinear regression on the log-transformed decline models with log-transformed production rate data gives better performance from the NLS algorithm (versus not using log transformations). Here, the Levenberg-Marquardt fitting algorithm gives relatively robust results, conditional on the convergence tolerance (˜1.5e-8) and maximum number of iterations allowed (1000). This fitting procedure is implemented by the *nlsLM* function in R (R Core Team, 2017).

Starting values for the NLS decline curve fits are either chosen as typical values from literature, estimated from the data, or determined pragmatically through trial and error. The Exponential model does not require starting values because it is linear. The Hyperbolic/Harmonic model uses starting values for q_i as the first production rate in the data series $(q_{t=1})$, $b = 0.5$, and

$$D_i = -\frac{\ln(q_2) - \ln(q_1)}{t_2 - t_1}. \tag{18}$$

The Power Law Loss-ratio model uses the same q_i and D_i starting values as above, and initial $D_\infty = 1e-6$ and $n = 0.15$. The Stretched Exponential also uses the first data point for the starting value of q_i and $n = 0.15$, but with $\tau = 2$. The Logistic Growth models uses the average parameter values reported in Clark et al. (2011): $K = 1.78e6$, $a = 33$, and $n = 0.9$. The starting values for Duong's model are $a = 0.25$, $m = 1.5$, and $q_\infty = 0$, with the same q_i as the previous models. Furthermore, all parameters are constrained to be non-negative in the NLS fitting algorithm. The GNLS algorithm is given the NLS-estimated parameter values as starting points.

In steps 3 and 5 above, the orders p and q of the ARMA models are determined by iteratively fitting models of varying order and taking the one with the best

Akaike information criterion (AIC):

$$AIC = 2d - 2\ln(\hat{L}), \tag{19}$$

where d is the number of parameters in the model and $\hat{L}$ is the maximum of the likelihood function for the model. AIC quantifies the goodness-of-fit of the model to the data in the likelihood function, but also adds a penalty for the number of parameters in order to prevent overfitting. Thus, in steps 3 and 5, we seek to find

$$\begin{aligned}[\hat{p}, \hat{q}] = &\arg\min_{p,q\in[0,1,\ldots,5]} AIC(\ln[f(t;\theta)] + \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + \\ &\theta_1 w_{t-1} + \theta_2 w_{t-2} + \ldots + \theta_q w_{t-q}).\end{aligned} \tag{20}$$

This optimization is performed with the *auto.arima* function in R (R Core Team, 2017).

2.1. Prediction Intervals.

Prediction intervals are determined via a bootstrap approach that starts by drawing 1000 random samples of decline curve parameters based on their mean and standard errors returned from the NLS or GNLS fits. The parameter standard errors are determined as the residual standard error times the square root of the Hessian matrix of the model's log likelihood function, which is estimated numerically (Ritz and Streibig, 2008, p.13). In the bootstrapping, all parameters are treated as independent and log-normally distributed (to prevent negative values and, thus, unrealistic decline curves). The decline model then predicts 1000 forecasts over the testing time period using these 1000 sets of parameter values in turn. The standard error at each forecasting time is computed as the square root of the sample variance of the 1000 log-transformed predictions ($\hat{\mathbf{q}}_\mathbf{t}$) at that time divided by the square root of the number of training samples:

$$SE_t = \sqrt{\frac{Var(\ln \hat{\mathbf{q}}_\mathbf{t})}{T_{train}}}. \tag{21}$$

The prediciton intervals are then constructed in log space as

$$PI_{t,\alpha} = [\ln \hat{q}_t - \frac{Qt_{\alpha/2,df}}{Q\mathcal{N}_{\alpha/2}} SE_t, \ln \hat{q}_t + \frac{Qt_{1-\alpha/2,df}}{Q\mathcal{N}_{1-\alpha/2}} SE_t], \tag{22}$$

where α is the significance level (in this paper, we use $\alpha = 0.2$ in order to give an 80% prediction interval, whose limits correspond to P90 and P10), $\hat{q}_t$ is the prediction at time t given by the mean decline curve parameter values, Qt is the quantile from a t-distribution with degrees of freedom $df = T_{train} - d$ (with d being the number of model parameters), and $Q\mathcal{N}$ is the quantile from a standard

Normal distribution. Note that because the GNLS approach estimates decline curve parameters with greater efficiency, and thus smaller standard error, one would expect the prediction intervals to be narrower than those from the NLS-fit models.

The above construction works for the NLS- and GNLS-fit decline curves; to get prediction intervals for their counterparts that include the ARMA-modeled residuals, we use the same equation above for $PI_{t,\alpha}$, but the prediction standard error is instead

$$SE_t = \sqrt{\frac{Var(\ln \hat{\mathbf{q}}_\mathbf{t}) + Var(\hat{\mathbf{e}}_\mathbf{t})}{T_{train}}}, \tag{23}$$

with $Var(\hat{\mathbf{e}}_\mathbf{t})$ being the variance of the predicted residuals from the ARMA model. This prediction variance is given via Kalman filter forecasting as described in Harvey and McKenzie (1982) and implemented in the R function *predict.Arima* (R Core Team, 2017). One can see that the prediction standard error that includes the ARMA model (Eq. 23) will always be at least as large as that for the base decline curve model (Eq. 21). That is, by capturing the behavior of past decline curve model residuals, the prediction intervals will tend to be wider, and thus more conservative.

3. Data

The West Virginia Geologic and Economic Survey and DrillingInfo supply monthly production histories, along with various metadata, for wells in the Marcellus formation. Out of an initial population of 15,990 wells, only 8,527 have suitable data for decline curve analysis. This subset was determined after filtering and cleaning the production histories to remove data before the stated completion dates, removing probable partial observations (months where not all days exhibited production), removing zeros, and keeping only a continuous record of gas rates until the first (if any) gap in the record. Furthermore, after all these pre-processing steps, only records with more than 24 months of data were retained, in order to have a sufficient number of data points for robust curve fitting and to retain some testing data for assessing forecasting accuracy.

4. Results and Discussion

4.1. Example with Duong's Model.

We start with a detailed illustration focusing on Duong's model exercised over all available Marcellus shale gas wells. First, a comparison is made between two methods of fitting Duong's decline curve model: the prescribed step-wise ordinary least squares (OLS) procedure (as outlined in Duong (2011)) and a non-linear least squares (NLS) procedure (this initial analysis is unique to Duong's

model). Fig. 1 compares the sum of squared residuals (SSR) from these two approaches, as fit to the 8,527 Marcellus records. For any one record, the better fitting model will give the smaller SSR, defined in Eq. 12. The NLS approach generally gives lower SSR than the OLS approach, with some exceptions (the vast majority of points fall below the 1:1 line in Fig. 1). This indicates that NLS fits Duong's decline model to the data better than the standard OLS approach.

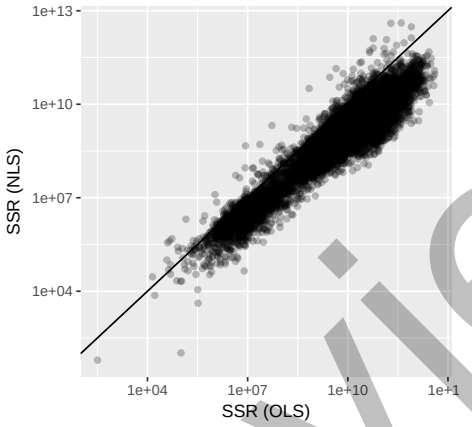

Figure 1: Sum of squared residuals from the non-linear least squares fitting approach versus sum of squared residuals from the traditional step-wise OLS approach. The diagonal line is the 1:1 ($x = y$) line.

Furthermore, because it fits the data better, the NLS fitting method yields less serial autocorrelation in the residuals, e. The Durbin-Watson test statistic serves as a metric for the significance of sample autocorrelation values (in this case, we only look at lag-1 autocorrelation, or the correlation between consecutive observations in time):

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}. \tag{24}$$

The value for DW is between 0 and 4, with smaller values indicating increasing positive autocorrelation and larger values indicating increasing negative autocorrelation ($DW = 2$ indicates no autocorrelation). Fig. 2 shows that the NLS fitting approach generally gives larger DW values than the OLS approach, again with some exceptions (majority of points fall above the 1:1 line, with few below). Thus, the bias introduced through the comparatively poor fits of the OLS approach gives greater positive autocorrelation in the residuals.

Although the NLS fitting approach reduces the magnitude of serial autocorrelation in the decline curve residuals, there is still considerable autocorrelation in a significant proportion of the well production histories. In Fig. 2, 2,946 points fall below $DW = 1$, which is generally accepted as critically low, for the NLS

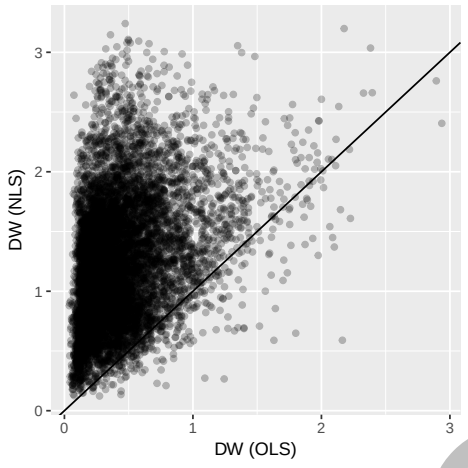

Figure 2: Durbin-Watson test statistics for the lag-1 autocorrelation value from the non-linear least squares fitting approach versus the traditional step-wise OLS approach. The diagonal line is the 1:1 ($x = y$) line.

approach (or approximately 35% of the 8,527 wells).

With the NLS regression established as the superior fitting procedure in the baseline case (over OLS; again, this is strictly for Duong's model), we can proceed to investigate the improvements afforded by GNLS regression and ARMA modeling of the residuals. For the rest of the analysis, the first $T_{train} = 24$ months of every well production history is used for model fitting, while the remainder of the record is reserved for validation (calculation of mean absolute percent error and coverage rate below). The next step is to fit an ARMA model to the calculated residuals from the NLS regression of Duong's model. Fig. 3 (subplot labeled "Duo") shows the distributions of autoregressive order p and moving average order q from fitting ARMA models to the Duong model residuals. While the case of $p = 0$ and $q = 0$ is predominant, there are cumulatively more cases with $p > 0$ and/or $q > 1$. That is, more often than not, significant autocorrelation is observed and an ARMA component is warranted for inclusion with Duong's model.

GNLS regression of Duong's model is run on all production records with any non-zero p or q value (if both $p = 0$ and $q = 0$, then there is no serial autocorrelation structure to specify and GNLS regression reduces to a weighted least squares regression). Separately, the fitted ARMA model from the OLS/NLS residuals is added to the OLS/NLS decline curve fit (as in Eq. 14-17). Fig. 4 shows an example of these different fitting procedures for Duong's decline curve model, where those methods incorporating ARMA-modeled residuals not only fit the training data better ($\leq$ 24 months), but also give more reliable prediction intervals after 24 months. The most important improvement here is the increase

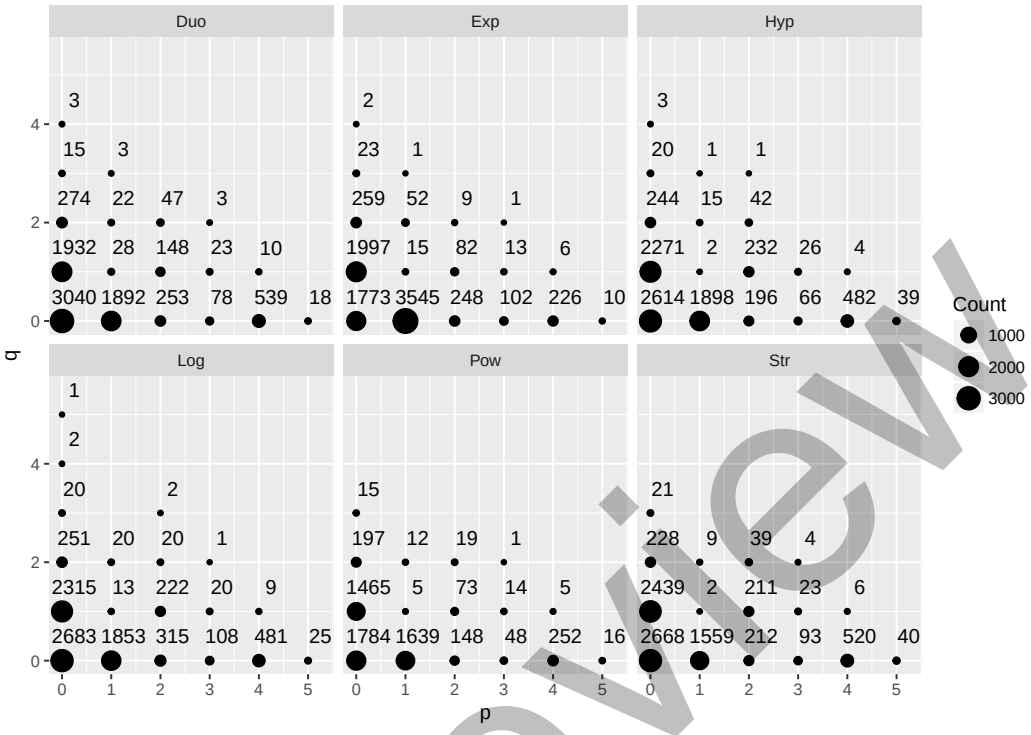

Figure 3: Plots showing distribution of AR and MA orders (p and q, respectively) for each decline curve model. The numbers in each plot indicate the number of wells with the associated combination of p and q values.

in CR afforded by the ARMA modeling, where the prediction intervals without ARMA fail to capture nearly all future data points. The ARMA modeling, by increasing the standard error, widens the prediction intervals considerably to capture more data points. The forecast accuracy remains about the same (MAPE $\approx 79\%$), where the slightly smaller value for the ARMA cases is driven by a short-term effect of the AR and MA components immediately after 24 months. This effect wears off, and the average prediction rebounds back to the same value as those cases without ARMA, in the long term. This particular case was chosen as it is one of the longer time series in the dataset (134 months), and gives a representative portrayal of what one would expect to see when estimating EUR with these various fitting approaches: no significant difference. Again, the real improvement comes when assessing uncertainty about the average EUR estimate, as the prediction intervals with the ARMA method scale with the error in the training data. Furthermore, in this example, the use of GNLS in fitting the models makes no impact on the goodness of fit, forecasting accuracy, or prediction

intervals.

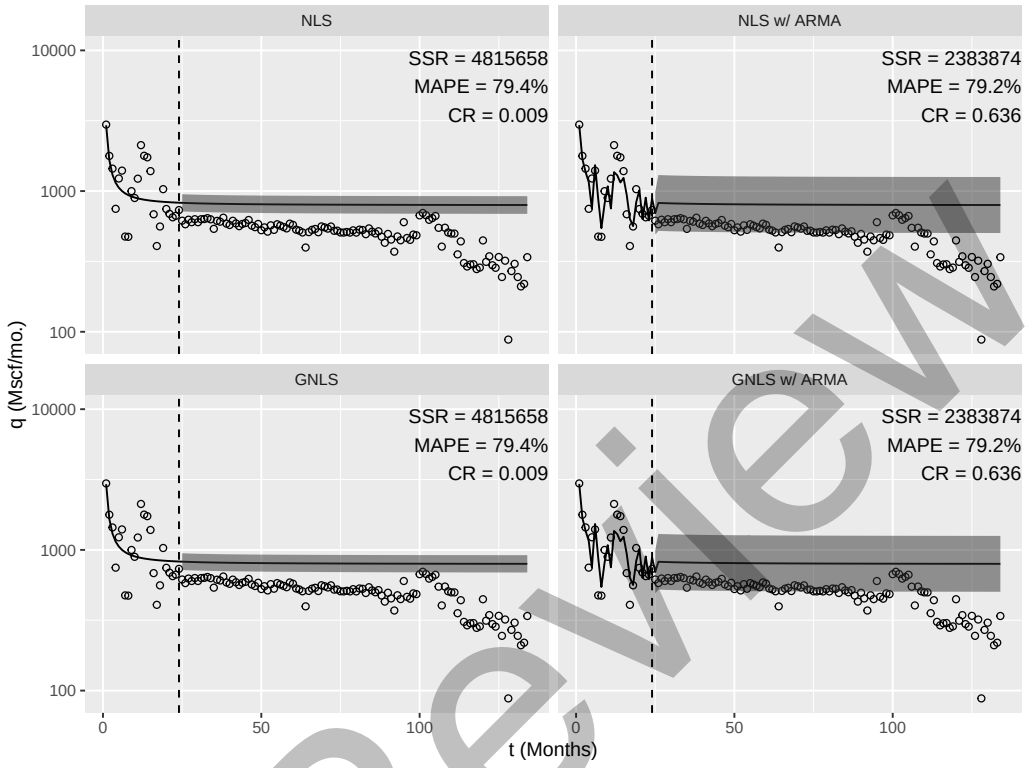

Figure 4: An example from well API 47-045-01807. The vertical dashed line is at $t = 24$ months, representing the cut-off between training and testing data points. The solid black curve is the estimated decline curve for each fitting procedure, and the shaded region around the curve after 24 months is the 80% prediction interval. Production data are provided by DrillingInfo.

The affect of these two modeling treatments (GNLS and ARMA) is measured by three performance metrics: 1) SSR, which indicates the goodness-of-fit to the training data (first 24 months), 2) mean absolute percent error (MAPE), which assesses the predictive accuracy of the model forecasts on the reserved testing data (after 24 months), and 3) coverage rate (CR) of the forecast prediction intervals, which counts the proportion of testing data falling within a model's prediction interval. SSR is defined in Eq. 12, but now the summation only goes to T_{train} instead of T. MAPE is calculated on the testing data as

$$MAPE = \frac{1}{T - T_{train}} \sum_{t=T_{train}+1}^{T} 100\% \left| \frac{q_t - \hat{q}_t}{q_t} \right|. \tag{25}$$

The coverage rate uses the prediction intervals at the $\alpha = 0.2$ significance level

as defined in the Prediction Intervals subsection above, and is calculated as the proportion of testing data points lying in the prediction interval. Ideally, CR = 0.8; that is, we expect 80% of the data to fall in between P90 and P10. In order to more easily compare favorable and unfavorable coverage rates (especially in the paired t-tests below), an adjusted CR metric

$$CR_{adj} = \frac{|CR - 0.8|}{0.8} \tag{26}$$

transforms CR such that good coverage rates (closer to 0.8) have low $\mathrm{CR_{adj}}$ values and poor coverage rate (far from 0.8) have high $\mathrm{CR_{adj}}$ values, on the [0, 1] scale.

Fig. 5) shows the SSR, MAPE, and $\mathrm{CR_{adj}}$ values for all the wells fit with Duong's model plotted by whether they incorporate ARMA-modeled residuals or not ("NLS w/ ARMA" versus "NLS w/o ARMA"), by whether they use GNLS or not ("GNLS" versus "NLS"), and the combined effect of GNLS and ARMA. In this plot, the salient points made from Fig. 4 are generally true when looking at all wells in the dataset, although to vary degrees and with many exceptions. Here we see that including the ARMA-model generally gives a better fit (lower SSR), a better forecast (lower MAPE), and a better coverage rate (lower $\mathrm{CR_{adj}}$), whereas the use of GNLS alone makes little apparent difference, except in the case of $\mathrm{CR_{adj}}$ where it can either improve or enhance the coverage rate, sometimes drastically so (many points dispersed away from the 1:1 line in Fig. 5h). The ARMA component largely drives the improvements in SSR and MAPE, where the patterns seen in the c.) and f.) subplots mimic the patterns in a.) and d.). Moreover, the improvement in MAPE that inclusion of the ARMA model affords appears to be primarily for NLS fits with relatively smaller prediction error to begin with ($MAPE \lessapprox 1000\%$), as seen in Fig. 5d. Only several points diverge significantly below the 1:1 line at higher MAPE values because those cases have production behavior that is too erratic after 24 months. Furthermore, in terms of coverage rate, the inclusion of the ARMA component can degrade performance if the NLS w/o ARMA model already has a good coverage rate close to 80% (see cluster of points above the 1:1 line at $CR_{adj} < 0.25$ in Fig. 5g). This tends to occur when the widening of the prediction interval (afforded by the addition of ARMA prediction standard error) leads to more than 80% of the testing points falling within the interval. Note that above $CR_{adj} > 0.25$, there is little risk of the ARMA modeling to detrimentally effect the coverage rate performance (few points lie above the 1:1 line in this domain).

Paired t-tests assess the statistical significance of these patterns. Since each well gives performance metric values in all categories (w/ ARMA, w/o ARMA, GNLS, NLS), pairing by well and taking the difference of metric values at this level gives a more powerful hypothesis test. Furthermore, since we want to test whether the model fitting procedures are making an improvement, one-sided tests

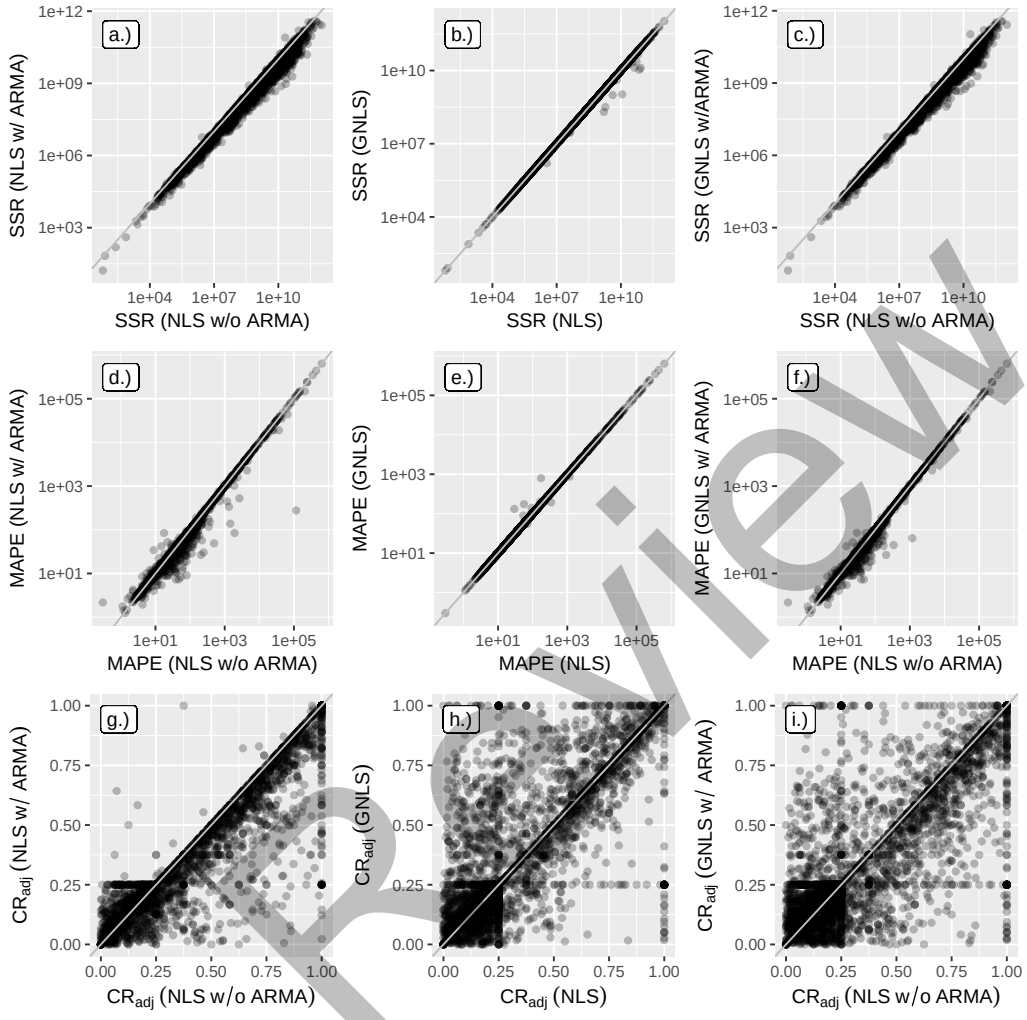

Figure 5: Plots showing performance metrics (SSR, MAPE, and CR_{adj}) from all wells fit with Duong's model and with $p > 0$ or $q > 0$. The performance metrics are plotted by the proposed fitting improvements (with ARMA, using GNLS, and the combination of the two for the regression) against the baseline case (no ARMA modeling of residuals and using NLS for the regression). The grey line in each plot is the 1:1 ($x = y$) line.

are appropriate. Specifically, with our null hypotheses being that there is no change in MAPE, SSR, or CR_{adj} with the inclusion of ARMA-modeled residuals, with using GNLS, or with the combination of ARMA and GNLS, our alternative hypotheses are that MAPE, SSR, and CR_{adj} are lower with these fitting methods than without ARMA and without GNLS. All tests are conducted at a 95% confidence level.

Table 1 presents the estimates of the mean of differences and their associated p-values from these tests. These values for Duong's model ("Duo" column) agree with the patterns seen in Fig. 5. SSR improves with all treatments (GNLS, with ARMA, and GNLS with ARMA), but the greatest improvement in this goodness of fit is with the inclusion of the ARMA component (reduces SSR by 2.8e9 $(\text{Mscf/month})^2$ on average). Similarly, the ARMA component makes the greatest reduction in MAPE on average: 51 percentage points (which is statistically significant). Fitting with GNLS tends to increase MAPE by 10 percentage points on average, and combining GNLS with ARMA gives moderate reduction of 19 percentage points (that is not statistically significant). Coverage rate is also most improved by the added ARMA model, but not to the extent that Fig. 5g suggests; ARMA only pushes the coverage 0.02 closer to the 0.8 expected value on average (however, this value is highly significant). As we see in Fig. 5g, there is certainly much greater potential for larger improvements (almost all the way up to 1.00), but these are counteracted by a large density of points on or very near to the 1:1 line and also the aforementioned points above the 1:1 line (for $\text{CR}_{\text{adj}} < 0.25$, especially).

4.2. Extension to Other Decline Curves.

The same comparative procedure made on Duong's decline curve model above is applied to five other decline curves: Arps' Exponential (Exp), Hyperbolic (Hyp), Power Law Loss Ratio (Pow), Stretched Exponential (Str), and Logistic Growth (Log). Again, the Exponential model uses OLS and GLS, whereas all other models use the nonlinear variants. The same steps outlined in the Methodology are applied to all 8,527 Marcellus wells for all decline curves. The frequency of orders p and q of ARMA models for the six different decline curves are shown in Fig. 3. Here, all other models show a similar distribution of p and q to Duong's model, with zero-order values being the predominant category (except with Exp), but with $p > 0$ and/or $q > 0$ cases cumulatively having a greater frequency.

Furthermore, the same one-sided paired t-tests as used for Duong's model above are run on the results from the other decline curves, and the results are presented in Table 1. In support of Table 1, and in lieu of the sort of scatterplots in Fig. 5, Fig. 6 shows boxplots of the raw MAPE, SSR, and CR_{adj} values, organized by decline curve type and fitting procedure. However, the main conclusions should be drawn from the t-tests (Table 1), because these boxplots do not pair the data on a well-by-well basis, which can be misleading. For example, in the Duo boxplots for CR_{adj}, we see no difference in the median values (these are all equal to 0.25 for all four boxplots), but the ARMA cases have lower third quartile (75th-percentile) values, which help explain the improvements in coverage rate for Duong's model reported in the t-testing above.

In Table 1, we see that including ARMA-modeled residuals always improves

	Decline Models					
Alternative	**Exp**	**Hyp**	**Pow**	**Str**	**Log**	**Duo**
SSR						
NLS > GNLS	-9.1e+10 (9.7e-01)	-6.9e+06 (9.5e-01)	**1.8e+06 (9.3e-02)**	-1.8e+06 (9.3e-01)	***3.4e+10 (6.2e-11)***	***5.7e+07 (6.4e-03)***
w/o ARMA > w/ ARMA	***3.0e+09 (1.0e-24)***	***2.2e+09 (2.6e-11)***	***2.8e+09 (2.6e-09)***	***1.8e+09 (3.6e-50)***	***2.5e+11 (7.0e-32)***	***2.8e+09 (3.9e-29)***
NLS w/o ARMA > GNLS w/ ARMA	***2.8e+09 (9.7e-90)***	***1.7e+09 (1.3e-58)***	***2.0e+09 (1.6e-43)***	***1.6e+09 (1.1e-52)***	***3.9e+10 (5.1e-13)***	***2.5e+09 (2.9e-27)***
MAPE						
NLS > GNLS	**1.7e+00 (4.2e-01)**	**6.6e-02 (1.3e-01)**	-5.8e+10 (8.4e-01)	**6.1e-03 (3.1e-01)**	***1.6e-01 (1.1e-02)***	-1.0e-01 (7.7e-01)
w/o ARMA > w/ ARMA	-2.7e+00 (8.1e-01)	**3.2e+00 (9.7e-02)**	-4.5e-01 (6.0e-01)	**7.5e-02 (4.4e-01)**	**5.6e+00 (8.3e-02)**	***5.1e+01 (2.2e-02)***
NLS w/o ARMA > GNLS w/ ARMA	**1.6e+00 (4.2e-01)**	**7.6e-01 (1.7e-01)**	-8.0e+10 (8.4e-01)	-2.3e-02 (5.2e-01)	**7.6e-01 (9.1e-02)**	**1.9e+01 (5.6e-02)**
CR_adj						
NLS > GNLS	***1.7e-01 (0.0e+00)***	***4.1e-03 (1.1e-02)***	**1.1e-03 (3.3e-01)**	**1.8e-03 (2.8e-01)**	-3.2e-03 (9.7e-01)	-1.9e-02 (1.0e+00)
w/o ARMA > w/ ARMA	***1.9e-01 (0.0e+00)***	***1.1e-01 (0.0e+00)***	**6.8e-06 (4.5e-01)**	***4.7e-02 (8.8e-231)***	***1.1e-01 (0.0e+00)***	***2.9e-02 (2.6e-94)***
NLS w/o ARMA > GNLS w/ ARMA	***2.9e-01 (0.0e+00)***	***1.1e-01 (0.0e+00)***	**1.1e-03 (3.4e-01)**	***5.1e-02 (4.7e-53)***	***9.7e-02 (6.0e-321)***	***1.2e-02 (1.1e-06)***

Table 1: Table of one-sided paired t-test results for SSR, MAPE, and CR_{adj} calculated over all decline models and at all viable well records, for 24 months of training data. In each cell, the number on top is the estimate of the mean of differences and the number in parenthesis is the p-value. Positive mean values are in bold and statistically significant values (p-value < 0.05) are italicized.

the fit to the training data (all models have positive estimates of mean of differences for SSR for the alternative hypothesis that "w/o ARMA > w/ ARMA", and these estimates are all statistically significant, some drastically so). Fitting with GNLS only helps in the Pow, Log, and Duo cases, with only the latter two being statistically significant; GNLS tends to worsen the goodness of fit for the Exp, Hyp, and Str decline models. Furthermore, while the combination of GNLS and ARMA also improves the fit for all decline models, it never does so to the same or higher degree as ARMA alone.

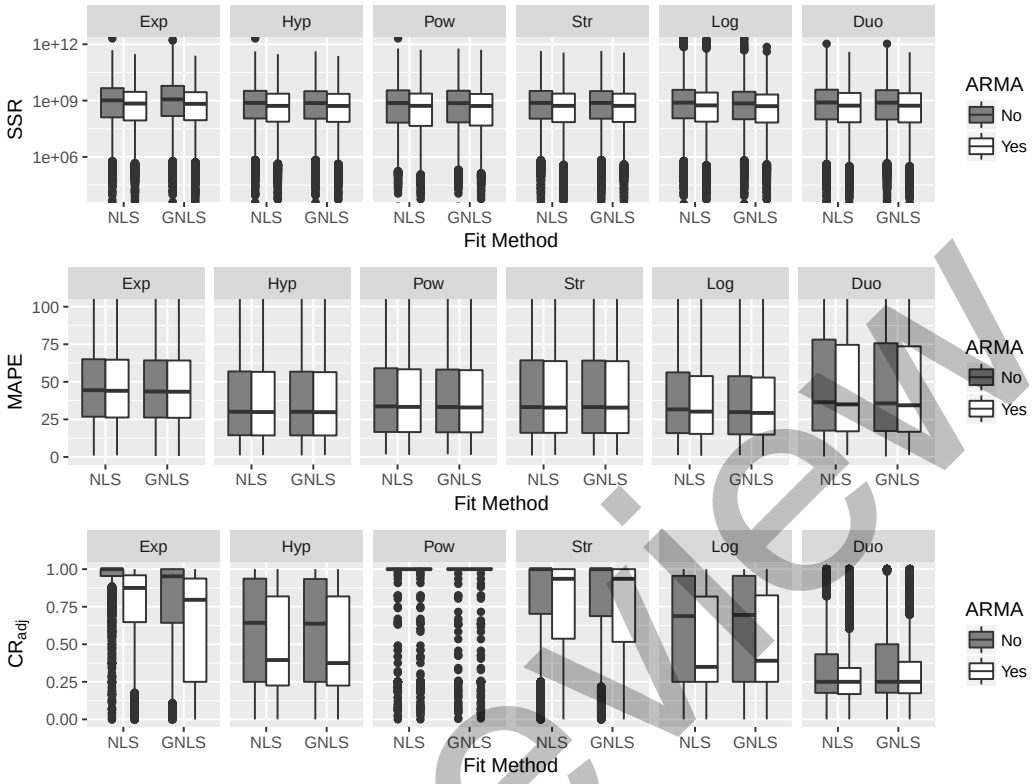

Figure 6: Box plots of SSR, MAPE, and CR_{adj} values for all viable Marcellus production histories after fitting each decline curve model under the various combinations of NLS versus GNLS and with ARMA-modeled residuals and without. The y axes have been clipped from 1e4 to 1e12 for SSR and from 0 to 100 for MAPE in order to exclude outliers and better visualize trends in the interquartile regions of the boxplots.

All but the Exp and Hyp models gain benefits to their forecasting ability (in terms of MAPE), with the only statistically significant instances being with Duong's model coupled with the ARMA model (as discussed above; note that this has the greatest improvement over all cases) and the Logistic Growth model when fit with GNLS (however only by 0.16 percentage points). Note that the Logistic Growth model with ARMA case reduces MAPE by 5.6 percentage points on average, but this is not statistically significant. While Duong's model shows the greatest improvement, the best overall case with respect to MAPE is the Hyperbolic model with ARMA-modeled residuals, which has the lowest median MAPE at 29.8% (Fig. 6; I gauge the lowest MAPE by median here because the distribution of MAPE is not normally distributed; the paired t-tests are still valid however, because they only require the differences of MAPE values to be

approximately normally distributed, which they are in this case).

The coverage rates generally improve significantly with use of the ARMA model (both with NLS and GNLS fits), except for the Power Law Loss Ratio model, which generates very poor prediction intervals at most wells (Fig. 6). As with Duong's model, these tend to be fairly modest gains on average, with the largest improvements occuring with the Exponential model (0.28 improvement in the rate when using GNLS with ARMA). However, it should be noted that Duong's model has the least to gain with respect to coverage rate, since the values for the NLS without ARMA case are already the lowest out of all decline curve models (Fig. 6); again, including ARMA on top of the NLS-fit Duong's model gives slightly better prediction intervals.

4.3. Extension to Longer Training Datasets.

The analysis above on all six candidate decline curves is extended to all wells with $T_{train} = 36$ months of training data (7,573 wells) and, subsequently, all wells with $T_{train} = 60$ months of training data (4,879 wells). The purpose here is not only to examine the effects of using more data, but more so to see how the patterns observed with $T_{train} = 24$ months may change with the potential inclusion of (more) data from boundary dominated flow conditions. The same paired t-tests are conducted and presented in Tables 2 (36 months) and 3 (60 months). The 36-month results (Table 2) have a similar pattern to that of 24 months (Table 1), except the major difference is that with 36 months, the MAPE consistently improves with the inclusion of the ARMA model across all six decline curves. All mean values are positive here, albeit only Duong's model remains statistically significant (this is also true for the GNLS with ARMA cases), with 62 percentage points improvement.

Also, the coverage rates generally improve across the board in the 36 month results as compared to the 24 month counterparts; all decline curves show larger mean values for both ARMA and GNLS with ARMA, and also for some GNLS cases. The largest improvements in the coverage of the prediction intervals come with the Exponential, Hyperbolic, and Logistic Growth models when they include the ARMA-modeled residuals and also with the "GNLS w/ ARMA" case; these improvements are all around a 0.2 improvement in the rate.

Looking at 60 months (Table 3) in comparison to 36 months, the mean differences for MAPE in the "w/o ARMA > w/ARMA" row are smaller for all decline curves even though all but the Exponential case are statistically significant. This suggests that at 60 months, the use of ARMA-modeled residuals with the NLS fits gives more consistency in making better predictions, at the sacrifice of magnitude of improvement. While half of the mean differences for MAPE in the "NLS > GNLS" row are now statistically significant, these improvements are not important in practice (0.3, 0.048 and 0.79 percentage points improvement). Sim-

	Decline Models					
Alternative	**Exp**	**Hyp**	**Pow**	**Str**	**Log**	**Duo**
SSR						
NLS > GNLS	-1.6e+15 (8.5e-01)	***1.3e+07 (4.9e-02)***	**8.6e+07 (1.7e-01)**	-1.6e+06 (5.6e-01)	***1.1e+11 (1.0e-11)***	**2.2e+08 (5.3e-02)**
w/o ARMA > w/ ARMA	***4.3e+09 (3.8e-88)***	***2.4e+09 (1.4e-57)***	***3.2e+09 (3.3e-34)***	***2.7e+09 (6.3e-22)***	***4.2e+11 (2.2e-34)***	***2.8e+09 (4.8e-11)***
NLS w/o ARMA > GNLS w/ ARMA	***4.5e+09 (1.5e-82)***	***2.1e+09 (8.8e-59)***	***2.6e+09 (7.6e-35)***	***2.6e+09 (1.4e-18)***	***1.1e+11 (3.2e-12)***	***2.0e+09 (7.7e-19)***
MAPE						
NLS > GNLS	-3.0e+00 (6.1e-01)	**9.3e-02 (2.9e-01)**	**1.0e+00 (3.5e-01)**	**2.9e-03 (3.5e-01)**	**1.7e+01 (1.6e-01)**	-2.2e-01 (7.8e-01)
w/o ARMA > w/ ARMA	**1.6e+01 (1.5e-01)**	**2.2e+01 (8.8e-02)**	**3.3e+00 (9.4e-02)**	**2.6e+00 (8.8e-02)**	**2.4e+01 (1.1e-01)**	***6.2e+01 (2.5e-02)***
NLS w/o ARMA > GNLS w/ ARMA	**1.1e+01 (2.8e-01)**	**2.7e+00 (1.1e-01)**	**6.2e-01 (3.7e-01)**	**2.2e+00 (1.3e-01)**	**2.7e+01 (1.1e-01)**	***5.7e+00 (2.1e-02)***
CR_adj						
NLS > GNLS	***1.7e-01 (0.0e+00)***	***2.7e-02 (3.9e-51)***	**3.3e-03 (1.4e-01)**	-4.9e-03 (9.3e-01)	***2.1e-02 (7.8e-31)***	**2.3e-03 (1.5e-01)**
w/o ARMA > w/ ARMA	***2.7e-01 (0.0e+00)***	***2.0e-01 (0.0e+00)***	-1.6e-04 (7.4e-01)	***8.1e-02 (1.9e-266)***	***2.0e-01 (0.0e+00)***	***5.8e-02 (8.6e-158)***
NLS w/o ARMA > GNLS w/ ARMA	***3.6e-01 (0.0e+00)***	***2.1e-01 (0.0e+00)***	**3.3e-03 (1.4e-01)**	***6.7e-02 (1.9e-65)***	***2.0e-01 (0.0e+00)***	***5.6e-02 (1.2e-90)***

Table 2: Table of one-sided paired t-test results for SSR, MAPE, and CR_{adj} calculated over all decline models and at all viable well records, for 36 months of training data. In each cell, the number on top is the estimate of the mean of differences and the number in parenthesis is the p-value. Positive mean values are in bold and statistically significant values (p-value < 0.05) are italicized.

ilarly, when assessing the GNLS with ARMA improvements, the gains in MAPE are small in comparison to just using ARMA alone (with the Exponential as an exception).

Again, the coverage rates improve with an even larger training data size (60 months). The Exponential, Hyperbolic, and Logistic Growth models remain the most improved when coupled with the ARMA model and also when fit with GNLS and coupled with the ARMA model. These improvements are all around 0.3, with the highest being 0.46 for the Exponential model under the "GNLS w/

	Decline Models					
Alternative	**Exp**	**Hyp**	**Pow**	**Str**	**Log**	**Duo**
SSR						
NLS > GNLS	-5.4e+09 (9.3e-01)	-6.2e+06 (7.2e-01)	***2.6e+08 (3.1e-02)***	**4.6e+06 (1.1e-01)**	***1.8e+11 (2.5e-09)***	-2.1e+09 (8.3e-01)
w/o ARMA > w/ ARMA	***5.7e+09 (5.2e-41)***	***2.3e+09 (7.8e-38)***	***3.7e+09 (7.5e-20)***	***2.6e+09 (1.9e-34)***	***4.8e+11 (2.2e-22)***	***2.2e+09 (1.6e-16)***
NLS w/o ARMA > GNLS w/ ARMA	***5.9e+09 (6.2e-36)***	***2.1e+09 (4.6e-40)***	***3.1e+09 (4.5e-17)***	***2.5e+09 (2.5e-31)***	***1.9e+11 (1.6e-09)***	***1.8e+09 (3.8e-18)***
MAPE						
NLS > GNLS	**6.9e+00 (8.8e-02)**	***3.0e-01 (2.5e-02)***	-4.2e-01 (9.7e-01)	***4.8e-02 (4.4e-03)***	***7.9e-01 (9.7e-09)***	**8.2e-02 (4.7e-01)**
w/o ARMA > w/ ARMA	**3.4e+00 (1.7e-01)**	***7.6e+00 (2.6e-02)***	***1.7e+00 (1.2e-02)***	***2.1e+00 (7.9e-03)***	***9.8e+00 (2.1e-02)***	***2.1e+01 (2.0e-02)***
NLS w/o ARMA > GNLS w/ ARMA	**7.1e+00 (9.2e-02)**	***3.1e+00 (9.3e-03)***	**9.5e-01 (1.3e-01)**	***2.1e+00 (1.3e-02)***	***3.1e+00 (1.4e-03)***	***5.4e+00 (5.6e-04)***
CR_adj						
NLS > GNLS	***2.5e-01 (2.7e-314)***	***5.1e-02 (7.4e-100)***	**2.8e-04 (4.7e-01)**	**2.9e-04 (4.8e-01)**	***3.9e-02 (2.7e-50)***	***1.6e-02 (1.2e-06)***
w/o ARMA > w/ ARMA	***4.1e-01 (0.0e+00)***	***3.2e-01 (0.0e+00)***	**7.8e-04 (7.5e-02)**	***1.2e-01 (1.1e-195)***	***3.0e-01 (0.0e+00)***	***1.2e-01 (7.6e-178)***
NLS w/o ARMA > GNLS w/ ARMA	***5.1e-01 (0.0e+00)***	***3.4e-01 (0.0e+00)***	**8.0e-04 (4.1e-01)**	***9.1e-02 (8.2e-50)***	***3.0e-01 (0.0e+00)***	***1.1e-01 (8.1e-99)***

Table 3: Table of one-sided paired t-test results for SSR, MAPE, and CR_{adj} calculated over all decline models and at all viable well records, for 60 months of training data. In each cell, the number on top is the estimate of the mean of differences and the number in parenthesis is the p-value. Positive mean values are in bold and statistically significant values (p-value < 0.05) are italicized.

ARMA" case.

4.4. Estimated Ultimate Recovery

One main purpose of DCA is to get Estimated Ultimate Recovery (EUR) values. EUR gives the overall gross worth of a producing well, and uncertainty in this value should always be reported along with the estimate. This uncertainty gives financial institutions and regulators a sense of the economic risk associated with continued operation of the well. Traditionally, one would integrate the

fitted decline curve (fit to all available data at a well) over a large time span, say 20 or more years, to obtain a P50 EUR estimate. Here, in order to test the ability of the prediction intervals in capturing EUR, a limited amount of initial data at each well is used to estimate EUR and the uncertainty in that estimate (prediction intervals, or P90 and P10 EUR values), while the remaining field data is used to calculate the actual observed EUR. Wells with only 10 years or more of data are used in this analysis, as the idea is to have some long-term measurements of EUR from the field upon which to evaluate the accuracy of the uncertainty quantification afforded by using ARMA-modeled residuals as compared to finding prediction intervals without the ARMA-modeling. 10 years was chosen as the lower threshold for well data duration because it is a sufficiently long time period and retains more wells in the dataset for testing (only one well had more than 20 years of data, while 220 wells have at least 10 years). It has already been demonstrated that the use of GNLS makes little difference, so that fitting procedure is not analyzed here. Also, the previous figures and tables show that the major improvements of using ARMA-modeled residuals lie in the coverage rate (i.e., the prediction intervals or uncertainty quantification), so the comparative MAPE of EUR estimates is not addressed here (fitting with or without ARMA-modeling performs about the same in this regard).

Specifically, the predicted EUR (P50) is found by fitting a decline curve to either the initial 24, 36, or 60 months of data at a well, and then integrating that fitted curve over the timespan from the end of the training data period to the last data point collected at the well and adding this integral to the sum of training data. Similarly, the uncertainty (P90 and P10 EUR values) is estimated by integrating the associated prediction intervals ($PI_{t,0.2}$; Eq. 22) over the same time frame and also adding these integrals to the cumulative training data. Our observed EUR values are taken as the cumulative of all available production data at each well with at least 10 years of production data. Figure 7 shows the proportion of 220 measured EUR values (from wells with more than 10 years of data) that fall within the P90-P10 bounds for different lengths of training data (24, 36, and 60 months) and the six different decline curve models examined in this paper.

The main result to be taken from Fig. 7 is that using ARMA-modeled residuals gives prediction intervals that always perform at least as well (and more often than not out-perform) as using prediction intervals without ARMA-modeling. That is, the white bars in Fig. 7 are usually higher than the grey bars, and seldom of equal height. This latter condition only occurs for all the Pow cases, which has already shown poor prediction interval performance (Fig. 6), and for the Str model with 24 months of training data.

Furthermore, the degree to which the estimates with ARMA-modeling out-perform those without tends to increase with the amount of training data (the

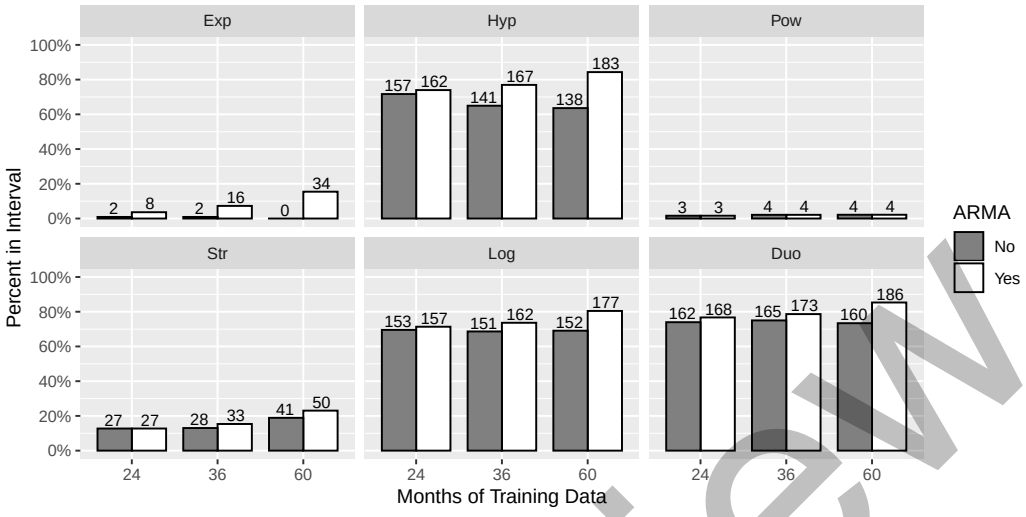

Figure 7: Barplots comparing the proportion of wells whose EUR values fall within the prediction intervals from fits using ARMA (white bars) versus fits without ARMA-modeled residuals (grey bars). This comparison is carried out for all six decline curve models examined in this paper, and for 24, 36, and 60 months of training data used for fitting the curves and estimating the prediction intervals. Numbers on top of each bar give the quantity of wells whose EUR fall within the prediction interval.

difference in height of the paired bars tends to grow with more months of data used for fitting). Including more data in the fitting process allows for a better characterization of the autocorrelation in the residuals, which leads to greater accuracy in estimating the prediction intervals (more specifically, in estimating the variance from the residuals, $Var(\hat{\mathbf{e}}_\mathbf{t})$). Note that at whichever wells the ARMA-modeling out-performs the regular methodology, the inclusion of the variance from the residuals is generating wider prediction intervals (larger difference between P90 and P10 EUR values).

To frame this in more practical terms, using ARMA-modeled residuals in DCA tends to give more accurate and more conservative estimation of the uncertainty about EUR at a well, with no possibility of doing worse than normal decline curve fitting. Because EUR is an important metric used by financial institutions and regulators to value oil and gas assets, the proposed method of using ARMA-modeling can help greatly in risk management and give a more accurate portrayal of financial liability to investors.

5. Conclusion

The main conclusion to be drawn from the analysis in this paper is that fitting an ARMA model to the residuals of a decline curve provides a rational and more accurate quantification of the uncertainty in production forecasts and estimates of ultimate recovery. The improved coverage rate of the prediction intervals and increased proportion of EUR values contained within P90 and P10 values illustrate the value of including the ARMA-predicted residual variance in the standard error term for the prediction intervals. This same treatment of including ARMA-modeled residuals consistently improves the forecasting accuracy (MAPE) of Duong's model to a statistically significant degree. The goodness-of-fit of the decline curve to historical data increases greatly when using the ARMA-modeled residuals, but this feature has little practical significance. Furthermore, estimating decline curve parameters via generalized (nonlinear) least squares as opposed to ordinary or nonlinear least squares does not appear to drive any of these improvements. A tangential conclusion made from the analysis in this paper is that Duong's model is generally fit better using nonlinear least squares than using the prescribed step-wise ordinary least squares procedure from Duong (2011).

Furthermore, in order to achieve a high level of statistical power, this study uses a BIG dataset of the entire population of Marcellus shale wells. Consequently, the fitting procedures in this paper were carried out in an automated fashion, where the algorithms used for the nonlinear least squares and generalized nonlinear least squares regressions are somewhat sensitive to the chosen starting values, convergence tolerances, and number of maximum iterations. While great effort was made in this analysis to find good heuristics for the starting values of each decline curve model, these could be suboptimal for some wells. Similarly, the global values used for convergence tolerance and maximum number of iterations may not have been appropriate for all wells. It is possible that different results may be achieved, and conclusions drawn, if one were to manually tailor these settings to each individual well. Nevertheless, the algorithms worked sufficiently for the majority of cases and I believe the conclusions are robust to the cases where they may have performed sub-optimally.

In terms of future work, more advanced time series modeling of the residuals (e.g., ARIMA modeling, GARCH modeling) may offer improvements beyond what is witnessed with the ARMA modeling in this study. Beyond a purely production data-driven modeling approach, it is worth exploring the possible dependency of autocorrelation in decline curve residuals on other exogenous variables, such as those related to the operation of the well. This may lead to more powerful forecasting models that including such information as explanatory variables, instead of simply modeling the residuals as a time series, as is done in this study. It may also suggest causal mechanisms for the (sometimes severe)

autocorrelation seen in many of the Marcellus wells in this analysis.

6. Acknowledgements

This work was made possible by the Deike Research Grant, College of Earth and Mineral Sciences, Penn State. The author thanks DrillingInfo for the donated academic license, as well as the journal editor for his valuable and constructive comments.

7. References

Arps, J.J., 1945. Analysis of Decline Curves. Transactions of the American Institute of Mining and Metallurgical Engineers 160, 228–247.

Ayeni, B.J., Pilat, R., 1992. Crude oil reserve estimation: An application of the autoregressive integrated moving average (ARIMA) model. Journal of Petroleum Science and Engineering 8, 13–28.

Cheng, Y., Wang, Y., McVay, D., Lee, W.J., 2010. Practical Application of a Probabilistic Approach to Estimate Reserves Using Production Decline Data. SPE Economics & Management 2, 19–31.

Clark, A.J., Lake, L.W., Patzek, T.W., 2011. Production Forecasting with Logistic Growth Models, in: SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers, Denver, Colorado, USA.

Cochrane, D., Orcutt, G.H., 1949. Application of Least Squares Regression to Relationships Containing Auto-Correlated Error Terms. Journal of the American Statistical Association 44, 32–61.

Duong, A.N., 1989. A New Approach for Decline-Curve Analysis, in: SPE Production Operations Symposium, Society of Petroleum Engineers.

Duong, A.N., 2011. Rate-Decline Analysis for Fracture-Dominated Shale Reservoirs. SPE Reservoir Evaluation & Engineering 14, 377–387.

Ediger, V.S., Akar, S., Ugurlu, B., 2006. Forecasting production of fossil fuel sources in Turkey using a comparative regression and ARIMA model. Energy Policy 34, 3836–3846.

Erdle, J., Hale, B., Houze, O., Ilk, D., Jenkins, C., Lee, W.J., Ritter, J., Seidle, J.P., Wilson, S., 2016. Monograph 4: Estimating Ultimate Recovery of Developed Wells in Low-Permeability Reservoirs. Society of Petroleum Evaluation Engineers.

Frausto-Solís, J., Chi-Chim, M., Sheremetov, L., 2015. Forecasting Oil Production Time Series with a Population-Based Simulated Annealing Method. Arabian Journal for Science and Engineering 40, 1081–1096.

Gallant, A.R., Goebel, J.J., 1976, 1976. Nonlinear regression with autocorrelated errors. Journal of the American Statistical Association 71, 961–967.

Gong, X., Gonzalez, R., McVay, D.A., Hart, J.D., 2014. Bayesian Probabilistic Decline-Curve Analysis Reliably Quantifies Uncertainty in Shale-Well-Production Forecasts. Spe Journal 19, 1,047–1,057.

Gupta, S., Fuehrer, F., Jeyachandra, B.C., 2014. Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis, in: SPE/CSUR Unconventional Resources Conference – Canada, Society of Petroleum Engineers.

Harvey, A.C., McKenzie, C.R., 1982. Finite-Sample Prediction From Arima Processes. Journal of the Royal Statistical Society Series C-Applied Statistics 31, 180–187.

de Holanda, R.W., Gildin, E., Valko, P.P., 2018. Combining Physics, Statistics, and Heuristics in the Decline-Curve Analysis of Large Data Sets In Unconventional Reservoirs. SPE Reservoir Evaluation & Engineering 21, 683–702.

Ilk, D., Rushing, J.A., Perego, A.D., Blasingame, T.A., 2008. Exponential vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps' Decline Curves, in: SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers.

Olominu, O., Sulaimon, A.A., 2014. Application of Time Series Analysis to Predict Reservoir Production Performance, in: SPE Nigeria Annual International Conference and Exhibition, Society of Petroleum Engineers.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Ritz, C., Streibig, J.C., 2008. Nonlinear Regression with R. Springer Science & Business Media, New York, NY.

Sheremetov, L., Cosultchi, A., Martinez-Munoz, J., Gonzalez-Sanchez, A., Jimenez-Aquino, M.A., 2014. Data-driven forecasting of naturally fractured reservoirs based on nonlinear autoregressive neural networks with exogenous input. Journal of Petroleum Science and Engineering 123, 106–119.

Shumway, R.H., Stoffer, D.S., 2010. Time Series Analysis and Its Applications. With R Examples, Springer.

SPE, 2018. Petroleum Resources Management System. Technical Report.

Valko, P.P., Lee, W.J., 2010. A Better Way To Forecast Production From Unconventional Gas Wells, in: SPE Annual Technical Conference and Exhibition, Society of Petroleum Engineers.

Wang, K., Li, H., Wang, J., Jiang, B., Bu, C., Zhang, Q., Luo, W., 2017. Predicting production and estimated ultimate recoveries for shale gas wells: A new methodology approach. Applied Energy 206, 1416–1431.

Weijermars, R., 2014. US shale gas production outlook based on well roll-out rate scenarios. Applied Energy 124, 283–297.

Weijermars, R., Sorek, N., Sen, D., Ayers, W.B., 2017. Eagle Ford Shale play economics: U.S. versus Mexico. Journal of Natural Gas Science and Engineering 38, 345–372.

Yuan, J., Luo, D., Feng, L., 2015. A review of the technical and economic evaluation techniques for shale gas development. Applied Energy 148, 49–65.

Yusof, N.M., Rashid, R.S.A., Mohamed, Z., 2010. Malaysia crude oil production estimation: an application of ARIMA model, in: 2010 International Conference on Science and Social Research (CSSR), IEEE. pp. 1255–1259.