PERBANDINGAN REGRESI LINEAR BERGANDA DAN *GENERALIZED ADDITIVE MODEL* UNTUK PREDIKSI *OIL RATE*

Muadz Aminuddin[1], Diyah Rosiani[1*], Pradini Rahalintar[1]
[1]Teknik Produksi Minyak dan Gas, Politeknik Energi dan Mineral Akamigas, Jalan Gajah Mada No. 38 Cepu, Blora, 58315
*E-mail: diyahrosiani@gmail.com

ABSTRAK

Analisis produksi sumur minyak memerlukan pemahaman tentang hubungan antara parameter operasional di permukaan dan laju alir minyak. Penelitian ini membandingkan kinerja regresi linier berganda dan *Generalized Additive Model* (GAM) dalam memprediksi *Oil Rate* (OR, Sm^3/hari) berdasarkan *Wellhead Pressure* (WHP), *Wellhead Temperature* (WHT), dan *Choke Size* (CS). Data berupa produksi harian sumur 15/9-F-12H Lapangan Volve periode 1 Januari–11 Agustus 2016 sebanyak 207 pengamatan (jam on-stream 24 jam). Korelasi Pearson menunjukkan CS berkorelasi positif terkuat terhadap OR (r = 0,58), diikuti oleh WHT (r = 0,32) dan WHP (r = −0,32), dengan multikolinearitas kuat antara WHP dan WHT (r = −0,92; VIF 7,16 dan 6,71). Pada pembagian acak 80:20 (42 data uji), R^2 uji regresi linier berganda sebesar 0,267, sedangkan GAM berbasis spline mencapai 0,543 dengan RMSE dan MAE lebih rendah. Namun, pada pembagian berbasis waktu dan *blocked cross-validation*, kedua model menghasilkan R^2 uji negatif karena OR menurun kuat seiring waktu (r = −0,95 terhadap hari produksi) dan bukaan choke diubah pada periode akhir sehingga data uji berada di luar rentang data latih. Dengan demikian, keunggulan GAM bersifat indikatif dan hanya terlihat pada interpolasi dalam periode yang sama, dan hubungan OR dengan WHP, WHT, dan CS tidak terpisah dari pengaruh waktu.

Kata kunci: ***Choke size, Generalized additive model, Oil rate,*** **Regresi linier berganda,** ***Wellhead pressure*****,** ***Wellhead temperature***

1. PENDAHULUAN

Prediksi laju alir minyak (*oil rate*, OR) merupakan aspek penting dalam manajemen produksi sumur minyak dan gas bumi. Laju alir yang melewati *choke* pada kepala sumur (*wellhead*) dipengaruhi oleh parameter operasional seperti *Wellhead Pressure* (WHP), *Wellhead Temperature* (WHT), dan *Choke Size* (CS). Berbagai korelasi empiris telah dikembangkan untuk memprediksi laju alir melalui *choke* berdasarkan parameter tersebut, mulai dari korelasi Gilbert [1] hingga model hibrida jaringan saraf tiruan dan pembelajaran mesin [2], [3], [4]. Model berbasis data juga makin banyak digunakan untuk prediksi produksi migas, mulai dari regresi linier hingga jaringan saraf dalam [5], [6], [7]. Studi-studi aliran melalui *choke* tersebut menunjukkan bahwa hubungan antara tekanan, ukuran *choke*, dan laju alir bersifat nonlinier [1], [2], [3], [4], sehingga regresi linier berganda, meskipun koefisiennya mudah diinterpretasikan, mengandung asumsi linieritas yang tidak selalu sesuai dengan aliran multifase pada sumur produksi [2]. Sebaliknya, model pembelajaran mesin kompleks seperti jaringan saraf tiruan atau *gradient boosting* umumnya akurat, tetapi kontribusi masing-masing prediktor sulit ditelusuri karena sifatnya yang *black box* [8].

Generalized Additive Model (GAM) menawarkan jalan tengah antara kedua pendekatan tersebut. GAM memodelkan setiap prediktor dengan fungsi mulus (*smooth function*) berbasis spline sehingga pola nonlinier dapat ditangkap tanpa menentukan bentuk fungsi di awal, sementara struktur aditifnya tetap mudah diinterpretasikan [9], [10], [11]. GAM telah diterapkan pada data produksi migas spasial-temporal lintas sumur dan lintas cekungan [10], tetapi per-

bandingan langsungnya dengan regresi linier berganda pada data operasional permukaan satu sumur (WHP, WHT, dan CS) masih jarang dibahas. Selain itu, korelasi yang tinggi antar-prediktor, khususnya antara WHP dan WHT, berpotensi menimbulkan multikolinearitas yang membuat koefisien regresi linier berganda tidak stabil dan sulit diinterpretasikan secara individual [12].

Kebaruan penelitian ini adalah perbandingan regresi linier berganda dan GAM yang dapat diinterpretasi pada level satu sumur. Kedua model dibangun pada gugus variabel permukaan yang sama (WHP, WHT, dan CS), dievaluasi dengan skema yang identik (pembagian acak, pembagian berbasis waktu, dan *blocked cross-validation*), dan ditafsirkan melalui koefisien regresi serta kurva *partial dependence* GAM. Berbeda dengan studi prediksi laju alir melalui *choke* yang menekankan akurasi model *black-box* [2], [4], [8], penelitian ini menekankan keterbacaan kontribusi tiap parameter sekaligus batas keberlakuan model ketika data bersifat deret waktu.

Penelitian ini bertujuan membandingkan kinerja regresi linier berganda dan GAM dalam memprediksi OR menggunakan data produksi harian sumur 15/9-F-12H di Lapangan Volve yang dirilis secara terbuka oleh Equinor [13] dan kerap digunakan dalam kajian *data-driven* di bidang teknik perminyakan [5]. Kinerja kedua model dinilai dengan koefisien determinasi (R^2), *Root Mean Square Error* (RMSE), dan *Mean Absolute Error* (MAE) pada data uji yang terpisah dari data latih [14].

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif berbasis data (*data-driven*) dengan data produksi harian sumur 15/9-F-12H, Lapangan Volve, Laut Utara Norwegia, yang dirilis secara terbuka oleh Equinor untuk keperluan penelitian dan pendidikan [13]. Data diambil dari lembar NO 15_9-F-12 H pada berkas Volve_production_2016_per_well.xlsx yang memuat 261 baris data harian periode 1 Januari–17 September 2016. Variabel bebas (X) adalah AVG_WHP_P (WHP, bar), AVG_WHT_P (WHT, °C), dan AVG_CHOKE_SIZE_P (CS, %), sedangkan variabel terikat (Y) adalah BORE_OIL_VOL (OR) yang dilaporkan dalam Sm^3/hari (1 $Sm^3 \approx$ 6,29 bbl). Lembar tersebut memuat nilai hilang, yaitu kolom BORE_WI_VOL kosong seluruhnya, AVG_DOWNHOLE_PRESSURE dan AVG_DOWNHOLE_TEMPERATURE bernilai 0 pada seluruh baris sehingga tidak digunakan, dan pada keempat kolom penelitian nilai hilang tidak berupa sel kosong melainkan tercatat sebagai nol saat sumur ditutup. Oleh karena itu, pemilihan data dilakukan dalam dua tahap: (i) 36 hari (13 Agustus–17 September 2016) dengan jam *on-stream* nol dikeluarkan karena OR = 0 serta WHP dan WHT tercatat 0 pada 35 hari (dan jauh di bawah kondisi operasi pada satu hari lainnya), sehingga bukan hasil pengukuran saat berproduksi; tersisa 225 hari; (ii) 18 hari dengan jam *on-stream* kurang dari 24 jam (ON_STREAM_HRS < 24) dikeluarkan karena volume harian pada hari tersebut tidak mencerminkan laju alir, dan rata-rata tekanan serta bukaan *choke* harian tidak representatif. Hasil akhirnya adalah 207 pengamatan harian (1 Januari–11 Agustus 2016) tanpa nilai hilang.

Tahapan analisis dimulai dengan statistik deskriptif (rata-rata, standar deviasi, minimum, median, dan maksimum), dilanjutkan dengan analisis korelasi Pearson beserta uji signifikansinya. Multikolinearitas antarvariabel bebas, yang dapat membuat koefisien regresi tidak stabil dan sulit diinterpretasikan secara individual, dideteksi dengan *Variance Inflation Factor* (VIF) dengan ambang VIF > 5 sebagai indikator yang perlu diwaspadai [12]:

$$VIF_{\mathrm{i}} = \frac{1}{(1 - R^2{}_{\mathrm{i}})} \tag{1}$$

dengan R^2_i adalah koefisien determinasi dari regresi variabel bebas X_i terhadap seluruh variabel bebas lainnya.

Evaluasi model dilakukan dengan tiga skema. Pertama, pembagian acak 80:20 (165 data latih dan 42 data uji) dengan *random_state* = 42 agar hasilnya *reproducible*. Kedua, karena OR harian merupakan deret waktu yang berautokorelasi dan menurun seiring umur sumur [6], [7], digunakan pembagian berbasis waktu: 80% pengamatan paling awal (166 hari; 1 Januari–24 Juni 2016) sebagai data latih dan 20% paling akhir (41 hari; 25 Juni–11 Agustus 2016) sebagai data uji. Ketiga, *blocked cross-validation* lima lipatan berurutan tanpa pengacakan. Pembagian acak menempatkan hari-hari yang bertetangga pada data latih dan data uji sehingga performa berpotensi overestimasi dan model dapat mempelajari proksi waktu alih-alih hubungan fisik; dua skema lainnya digunakan untuk menguji hal tersebut. Pada semua skema, parameter model hanya diestimasi dari data latih; untuk interpretasi koefisien dan kurva GAM (Subbab 3.B dan 3.C), model juga dibangun pada seluruh 207 data. Model pertama adalah regresi linier berganda yang diestimasi dengan metode *Ordinary Least Squares* (OLS):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{2}$$

dengan Y adalah OR; X_1, X_2, dan X_3 berturut-turut adalah WHP, WHT, dan CS; β_0 adalah konstanta; β_1, β_2, β_3 adalah koefisien regresi; dan ε adalah galat. Signifikansi setiap koefisien diuji dengan uji-t pada taraf 5%.

Model kedua adalah GAM, yang memodelkan OR sebagai penjumlahan fungsi mulus dari setiap prediktor tanpa mengasumsikan bentuk hubungan linier [9], [11]:

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_3) + \varepsilon \tag{3}$$

dengan f_1, f_2, dan f_3 adalah fungsi mulus berbasis spline kubik untuk WHP, WHT, dan CS, sedangkan simbol lainnya sama dengan Persamaan (2). Bentuk fungsi diestimasi langsung dari data, sehingga perubahan kemiringan pengaruh suatu variabel (misalnya pada rentang bukaan *choke* tertentu) dapat ditangkap tanpa ditentukan di awal [2], [3].

Model GAM dibangun dengan pustaka pyGAM [15] (versi 0.12.0, kelas LinearGAM) menggunakan fungsi mulus berbasis *penalized B-spline* kubik dengan 10 basis dan penalti turunan kedua pada setiap variabel. Parameter penalti (λ) dipilih hanya dari data latih melalui *grid search* dengan kriteria *Generalized Cross-Validation* (GCV) atas sembilan nilai berskala logaritmik pada rentang 10^{-3} hingga 10^5; satu nilai λ yang sama digunakan untuk ketiga fungsi. Derajat kebebasan efektif (EDF) setiap fungsi dilaporkan pada Tabel 4, dan R^2 data latih GAM dibandingkan dengan R^2 data uji untuk menilai *overfitting*.

Kinerja model pada data uji dinilai dengan R^2, RMSE, dan MAE (Persamaan (4)–(6)). R^2 menunjukkan proporsi variasi OR yang dijelaskan model, sedangkan RMSE dan MAE menyatakan besar galat prediksi dalam satuan OR (Sm^3/hari); ketiganya digunakan bersama karena kepekaannya terhadap galat berbeda [14]. Seluruh analisis dijalankan dengan Python, menggunakan pustaka statsmodels [16] untuk regresi linier berganda dan pyGAM [15] untuk GAM.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

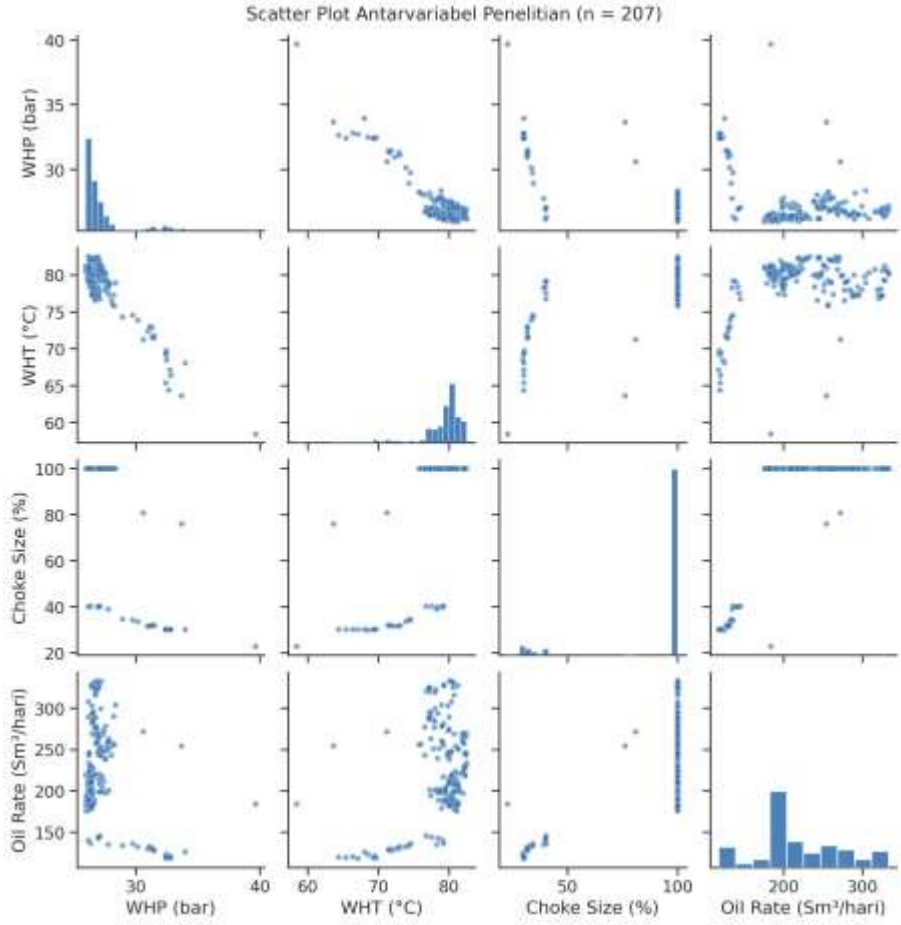
dengan y_i adalah nilai OR aktual ke-i, $\hat{y}_i$ adalah nilai OR prediksi ke-i, $\bar{y}$ adalah rata-rata OR aktual pada data uji, dan n adalah jumlah pengamatan pada data uji.

3. PEMBAHASAN

A. Statistik Deskriptif dan Analisis Korelasi

Statistik deskriptif dari 207 data pengamatan disajikan pada Tabel 1. Rata-rata OR pada sumur 15/9-F-12H selama periode pengamatan adalah 217,34 Sm³/hari dengan standar deviasi 53,89 Sm³/hari, yang mengindikasikan variasi produksi harian yang cukup besar. CS berkisar antara 22,91% dan 100,00% dengan median 100%, artinya 177 dari 207 pengamatan (85,5%) berada pada *choke* terbuka penuh, sedangkan WHP relatif stabil pada kisaran 26 hingga 27 bar dengan beberapa nilai ekstrem hingga 39,67 bar.

Gambar 1 menyajikan sebaran data antarvariabel dalam bentuk *scatter plot* berpasangan. Hubungan CS dan OR positif tetapi tidak linier, dan datanya terbagi menjadi dua kelompok: pada CS sekitar 30–41% OR hanya sekitar 118–145 Sm³/hari, sedangkan pada CS = 100% (85,5% pengamatan) OR menyebar lebar dari sekitar 176 hingga 333 Sm³/hari; hanya dua pengamatan berada di antara CS 41% dan 99%. Sebaran pada CS = 100% bukan efek jenuh (*saturation*) yang melandai, melainkan variasi OR yang tidak dapat dijelaskan oleh CS karena *choke* sudah terbuka penuh. Hubungan WHT terhadap OR menunjukkan pola serupa: OR meningkat pada WHT rendah hingga menengah, kemudian menyebar lebar pada WHT di atas sekitar 76°C.

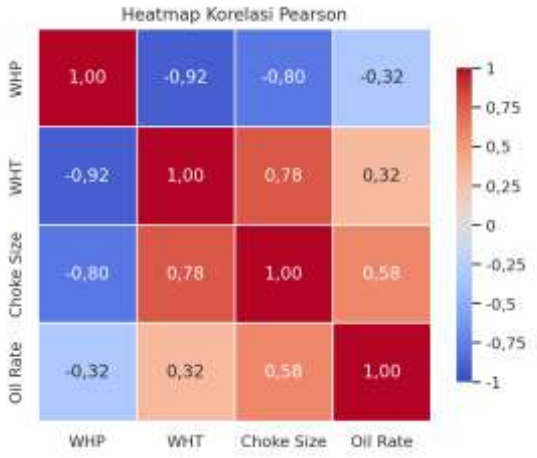


Gambar 1. *Scatter Plot* Antarvariabel Penelitian

Tabel 1. Statistik Deskriptif Variabel Penelitian (*Oil Rate* dalam Sm³/hari)

Statistik	WHP (bar)	WHT (°C)	*Choke Size*(%)	*Oil Rate* (Sm³/hari)
Rata-rata	27,19	78,85	91,72	217,34
Std. Deviasi	1,86	3,70	21,91	53,89
Minimum	25,96	58,45	22,91	118,12
Median	26,52	80,00	100,00	201,32
Maksimum	39,67	82,49	100,00	332,95

Analisis korelasi Pearson (Gambar 2) menunjukkan bahwa CS memiliki korelasi positif terkuat terhadap OR ($r = 0,58$; $p < 0,001$), diikuti oleh WHT ($r = 0,32$; $p < 0,001$) dan WHP yang berkorelasi negatif ($r = -0,32$; $p < 0,001$); seluruhnya signifikan pada taraf 5%. CS berpengaruh paling langsung karena mengatur bukaan aliran fluida menuju permukaan, sedangkan WHP yang tinggi umumnya terjadi ketika *choke* lebih tertutup sehingga laju alir lebih rendah.



Gambar 2. *Heatmap* Korelasi Antarvariabel Penelitian

Korelasi antarvariabel bebas juga sangat kuat, yaitu antara WHP dan WHT ($r = -0,92$), WHT dan CS ($r = 0,78$), serta WHP dan CS ($r = -0,80$). Kondisi ini mengindikasikan multikolinearitas yang dapat memperbesar variansi dan *standard error* koefisien regresi sehingga uji signifikansi koefisien individual kurang dapat diandalkan [12]. Secara fisik, WHP dan WHT sama-sama dipengaruhi oleh kondisi aliran: ketika *choke* makin terbuka, WHP cenderung turun sementara WHT cenderung naik akibat laju alir yang lebih besar.

Nilai VIF pada Tabel 2 mengonfirmasi indikasi tersebut: VIF WHP (7,16) dan WHT (6,71) melampaui ambang 5, sedangkan VIF CS (2,84) masih dalam batas wajar. Hal ini perlu diperhitungkan dalam interpretasi koefisien regresi linier berganda.

Tabel 2. Hasil Perhitungan *Variance Inflation Factor* (VIF)

Variabel	VIF
WHP	7,16
WHT	6,71
Choke Size	2,84

B. Analisis Regresi Linier Berganda

Untuk keperluan interpretasi, regresi linier berganda (OLS) dibangun pada seluruh 207 data; hasil estimasi koefisiennya disajikan pada Tabel 3.

Tabel 3. Hasil Estimasi Koefisien Regresi Linier Berganda

Variabel	Koefisien	Std. Error	Nilai-t	p-value
Konstanta	−100,50	263,49	−0,38	0,703
WHP (bar)	8,81	4,23	2,08	0,038
WHT (°C)	−1,60	2,06	−0,78	0,436
Choke Size(%)	2,23	0,23	9,88	< 0,001

CS merupakan variabel dengan koefisien paling signifikan ($p < 0,001$), konsisten dengan analisis korelasi. WHP signifikan pada taraf 5% ($p = 0,038$), sedangkan WHT tidak signifikan ($p = 0,436$), kemungkinan besar karena multikolinearitas dengan WHP sehingga kontribusi WHT sulit dipisahkan secara statistik [12]. Persamaan regresi yang terbentuk mengikuti Persamaan (2) adalah:

$$Oil\ Rate = -100,50 + 8,81(WHP) - 1,60(WHT) + 2,23(Choke\ Size) \quad (7)$$

Persamaan (7) menunjukkan bahwa setiap kenaikan 1% CS meningkatkan OR rata-rata sebesar 2,23 Sm^3/hari dan setiap kenaikan 1 bar WHP diasosiasikan dengan kenaikan OR sebesar 8,81 Sm^3/hari, dengan variabel lain tetap. Tanda positif koefisien WHP berbeda dari korelasi sederhananya yang negatif ($r = -0,32$). Perbedaan ini tidak semata-mata akibat multikolinearitas, tetapi juga konsisten dengan fisika aliran melalui *choke*: pada bukaan *choke* yang sama, tekanan hulu (WHP) yang lebih tinggi menghasilkan laju alir yang lebih besar, sebagaimana tercermin pada korelasi Gilbert [1]. Korelasi bivariatnya bernilai negatif karena WHP turun ketika *choke* dibuka lebih lebar (r WHP–CS = −0,80), sehingga pengaruh CS tercampur di dalamnya. Model ini menghasilkan $R^2 = 0,396$ pada seluruh data, artinya ketiga variabel secara linier hanya menjelaskan kurang dari separuh variasi OR, sejalan dengan sifat nonlinier hubungan tekanan, ukuran *choke*, dan laju alir [1], [2], [3].

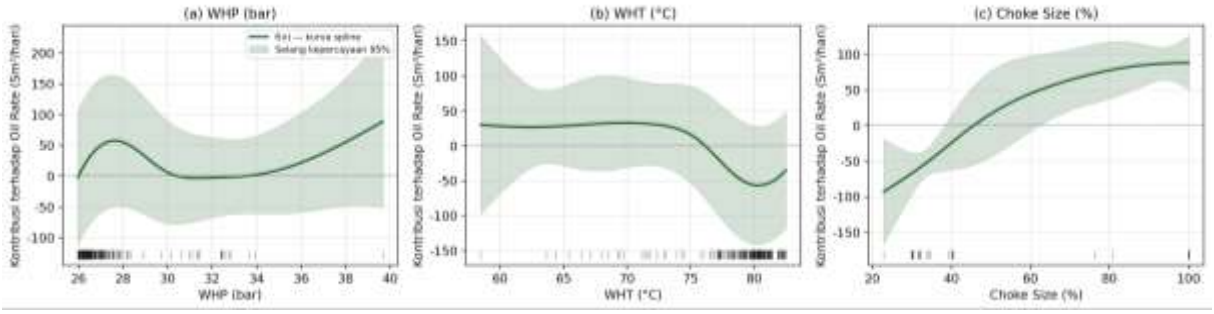
C. Analisis Generalized Additive Model (GAM)

GAM memodelkan OR sebagai penjumlahan fungsi spline kubik dari WHP, WHT, dan CS, sehingga setiap variabel dapat memiliki bentuk kurva pengaruh yang fleksibel dengan tetap mempertahankan interpretasi aditif per variabel [9], [11]. Pengaturan dan kompleksitas fungsi mulus disajikan pada Tabel 4. Pada seluruh data, GAM mencapai $R^2 = 0,585$ (regresi linier berganda 0,396). EDF terbesar dimiliki WHP (6,96) dan WHT (3,62), sedangkan CS relatif sederhana (1,28); EDF WHP yang tinggi pada data yang terkonsentrasi di 26–28 bar dapat menjadi tanda fleksibilitas yang berlebih.

Tabel 4. Pengaturan dan Derajat Kebebasan Efektif (EDF) Fungsi Mulus GAM (Model pada Seluruh Data)

Variabel	Jumlah basis	λ terpilih	EDF
WHP	10	0,1	6,96
WHT	10	0,1	3,62
Choke Size	10	0,1	1,28
Total model	–	–	11,86

Gambar 3 menyajikan kurva fungsi spline (*partial dependence*) ketiga prediktor beserta selang kepercayaan 95%; tanda vertikal di dasar panel menunjukkan sebaran data. Pada Gambar 3(a), kontribusi WHP tidak monoton: meningkat hingga sekitar +57 Sm^3/hari pada WHP ≈ 27,5 bar, turun mendekati nol pada 30–34 bar, lalu naik kembali di atas 34 bar. Selang kepercayaannya sangat lebar karena sebagian besar data berada pada 26–28 bar, dan liukan pada rentang tersebut dapat menjadi tanda penalti yang terlalu kecil (EDF = 6,96; λ = 0,1 dipakai bersama oleh ketiga fungsi), sehingga bentuk kurva WHP tidak dapat ditafsirkan terlalu rinci. Pada Gambar 3(b), kontribusi WHT relatif datar (sekitar +30 Sm^3/hari) pada 58–75°C, kemudian menurun hingga sekitar −55 Sm^3/hari pada WHT ≈ 80°C, tempat sebagian besar data terkonsentrasi (median 80,00°C); selang kepercayaannya mencakup nol pada sebagian besar rentang sehingga efek WHT tidak dapat dibedakan secara meyakinkan dari nol, dan lembah kurva di sekitar 80°C yang naik kembali menjelang 82°C juga tidak dapat ditafsirkan terlalu rinci. Gambar 3(c) memperlihatkan kontribusi CS yang meningkat dari sekitar −95 Sm^3/hari pada CS ≈ 23% menjadi sekitar +45 Sm^3/hari pada CS ≈ 60%, lalu melandai menuju sekitar +88 Sm^3/hari pada CS 80–100%. Bentuk melandai ini tidak dapat ditafsirkan sebagai efek jenuh fisik: karena hanya dua pengamatan berada di antara CS 41% dan 99%, kurva pada rentang tersebut terutama merupakan interpolasi antara kelompok CS rendah dan kelompok CS = 100%, sedangkan variasi OR pada CS = 100% (85,5% data) tetap tidak dapat dijelaskan oleh CS, sejalan dengan Gambar 1.

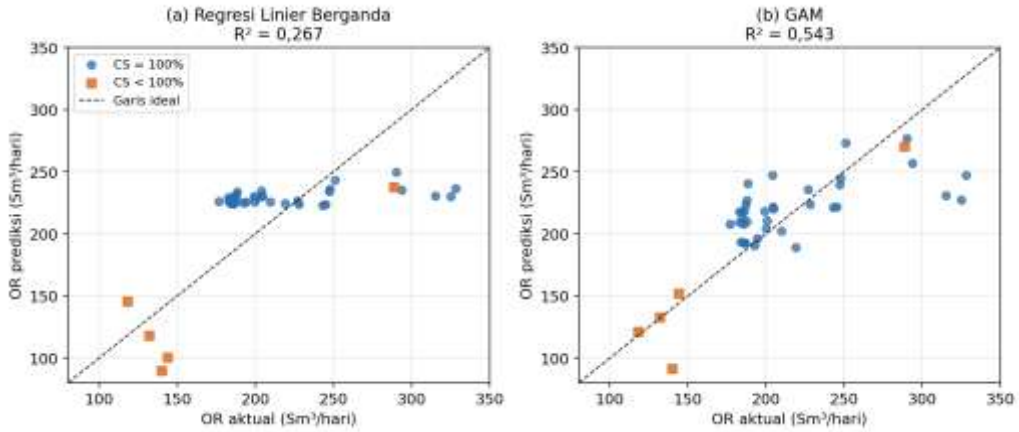


Gambar 3. Kurva Fungsi Spline (*Partial Dependence*) Model GAM untuk Variabel (a) WHP, (b) WHT, dan (c) *Choke Size*

Pola tersebut mendukung nonlinieritas yang teramati secara kualitatif pada Gambar 1, terutama pada CS. Untuk WHT, selang kepercayaan yang mencakup nol menunjukkan bahwa koefisien WHT yang tidak signifikan pada regresi linier (Tabel 3) tidak dapat dijelaskan semata-mata oleh bentuk nonlinier; multikolinearitas dengan WHP (VIF = 6,71) tetap menjadi penjelas utama. Selang kepercayaan juga melebar pada ujung-ujung rentang data, terutama pada WHP di atas 34 bar dan WHT di bawah 65°C, karena pengamatan pada wilayah tersebut jarang.

D. Perbandingan Kinerja Model

Kinerja kedua model dievaluasi dengan tiga skema sebagaimana diuraikan pada Metode. Gambar 4 menyajikan nilai aktual terhadap nilai prediksi pada 42 data uji pembagian acak; titik biru menandai hari dengan CS = 100% dan titik oranye menandai hari dengan CS < 100%.



Gambar 4. Perbandingan Nilai Aktual dan Prediksi (a) Regresi Linier Berganda dan (b) GAM

Prediksi GAM (Gambar 4b) tersebar lebih dekat ke garis ideal dibandingkan regresi linier berganda (Gambar 4a). Prediksi regresi linier berganda mendatar di sekitar 223–250 Sm³/hari untuk OR aktual 177–328 Sm³/hari, dan GAM pun tidak mampu mengikuti OR di atas sekitar 300 Sm³/hari. Pola ini berkaitan dengan dominasi *choke* terbuka penuh: 37 dari 42 data uji memiliki CS = 100%, sehingga variasi OR tidak dapat dijelaskan oleh CS, sedangkan WHP dan WHT pada kondisi tersebut hanya bervariasi dalam rentang sempit. Pada subset CS = 100% ini, R^2 regresi linier berganda hanya 0,037 (RMSE 41,95 Sm³/hari) dan R^2 GAM 0,368 (RMSE 33,99 Sm³/hari). Rekapitulasi metrik evaluasi disajikan pada Tabel 5.

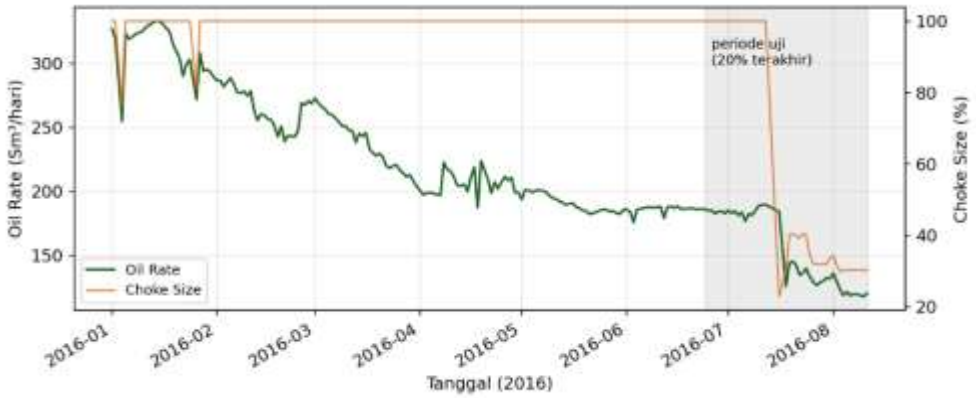
Tabel 5. Perbandingan Metrik Evaluasi Model pada Pembagian Acak 80:20 (42 Data Uji; RMSE dan MAE dalam Sm³/hari)

Metode	R^2 latih	R^2 uji	RMSE	MAE
Regresi linier berganda	0,418	0,267	41,70	36,28
GAM	0,589	0,543	32,92	24,26

Pada pembagian acak, GAM mengungguli regresi linier berganda pada ketiga metrik (Tabel 5). Selisih R^2 latih dan R^2 uji GAM (0,589 terhadap 0,543) lebih kecil daripada regresi linier berganda (0,418 terhadap 0,267), sehingga tidak tampak indikasi *overfitting* yang kuat pada GAM dengan λ hasil GCV. Namun, hasil ini dihitung hanya pada 42 data uji dan bergantung pada pembagian acak: pada 20 kunci acak berbeda, R^2 uji regresi linier berganda berkisar 0,096–0,515 (median 0,355) dan GAM 0,239–0,694 (median 0,560), dengan GAM lebih baik pada seluruh 20 pembagian.

Karena OR harian merupakan deret waktu, kinerja juga dievaluasi dengan pembagian berbasis waktu dan *blocked cross-validation* (Tabel 6). Pada pembagian berbasis waktu (data uji 25 Juni–11 Agustus 2016), kedua model gagal: R^2 uji regresi linier berganda −96,43 (RMSE 272,16; MAE 222,06 Sm³/hari) dan GAM −1467,48 (RMSE 1056,60; MAE 827,22 Sm³/hari), jauh lebih buruk daripada prediktor rata-rata data latih (RMSE 85,94 Sm³/hari). Penyebabnya terlihat pada Gambar 5: pada Januari–Juni, CS = 100% pada 167 dari 172 hari sementara OR turun dari sekitar 333 menjadi 176 Sm³/hari, sehingga penurunan tersebut tidak dapat dijelaskan oleh WHP, WHT, maupun CS. Pada Juli–Agustus, *choke* dikecilkan hingga 23–41%, WHP naik hingga 39,67 bar, dan WHT turun hingga 58,45°C, yaitu di luar rentang data latih (CS ≥ 76,1%; WHP ≤ 33,7 bar; WHT ≥ 63,6°C), sehingga model harus berekstrapolasi. Fungsi spline GAM tidak dirancang untuk ekstrapolasi sehingga galatnya jauh lebih besar daripada regresi

linier dan sensitif terhadap jumlah basis (R^2 uji −62,46; −1467,48; dan −12,49 untuk 6, 10, dan 15 basis).



Gambar 5. Deret Waktu *Oil Rate* dan *Choke Size* Sumur 15/9-F-12H (Area Abu-abu = Periode Uji pada Pembagian Berbasis Waktu)

Tabel 6. Kinerja Model pada *Blocked Cross-Validation* Lima Lipatan (RMSE dalam Sm³/hari)

Lipatan (periode; jumlah data uji)	RMSE regresi	RMSE GAM	RMSE rata-rata latih	R^2 regresi	R^2 GAM
1 (01/01–16/02; n = 42)	104,27	93,16	104,13	−16,67	−13,11
2 (17/02–30/03; n = 42)	35,33	29,49	37,82	−2,86	−1,69
3 (31/03–11/05; n = 41)	32,69	32,43	18,28	−15,18	−14,92
4 (12/05–24/06; n = 41)	53,28	34,27	39,20	−330,14	−135,96
5 (25/06–11/08; n = 41) = pembagian berbasis waktu	272,16	1056,60	85,94	−96,43	−1467,48
Rata-rata ± SD	99,55 ± 100,69	249,19 ± 452,13	57,07 ± 36,20	−92,26 ± 138,07	−326,63 ± 640,10

Pada *blocked cross-validation*, R^2 uji negatif pada seluruh lipatan karena variasi OR di dalam setiap lipatan kecil akibat tren penurunan (simpangan baku OR dalam lipatan 3,0–27,9 Sm³/hari, dibandingkan dengan 53,9 Sm³/hari pada seluruh data). GAM memiliki RMSE lebih rendah daripada regresi linier pada lipatan 1, 2, dan 4, hampir sama pada lipatan 3, dan jauh lebih tinggi pada lipatan 5; terhadap prediktor rata-rata data latih, GAM lebih baik pada lipatan 1, 2, dan 4, sedangkan regresi linier hanya pada lipatan 2. Dengan demikian, keunggulan GAM tidak konsisten antarperiode, dan rata-rata RMSE GAM (249,19 ± 452,13) lebih besar daripada regresi linier (99,55 ± 100,69), terutama akibat lipatan 5.

Keunggulan GAM pada pembagian acak sejalan dengan dugaan bahwa hubungan WHP, WHT, dan CS terhadap OR tidak sepenuhnya linier, serta dengan studi aliran melalui *choke* yang menunjukkan keunggulan model nonlinier dibandingkan korelasi linier sederhana [2], [3], [4] dan kajian tinjauan model berbasis data pada produksi migas [17], [18]. Akan tetapi, hasil pembagian berbasis waktu menunjukkan bahwa keunggulan tersebut tidak berlaku ketika model harus memprediksi periode di luar rentang data latih.

Nilai R^2 GAM sebesar 0,543 pada pembagian acak tergolong moderat, sehingga sekitar 46% variasi OR belum dapat dijelaskan. Diagnostik waktu menunjukkan alasan utamanya: OR berkorelasi sangat kuat dengan hari produksi (r = −0,95); hari produksi saja menjelaskan 90,4%

variasi OR, sedangkan ketiga variabel permukaan menjelaskan 39,6%, dan menambahkan ketiganya ke model berbasis hari hanya menaikkan R^2 menjadi 90,9%. Dengan demikian, sebagian besar kemampuan prediksi pada pembagian acak kemungkinan berasal dari keterkaitan ketiga variabel dengan waktu (proksi waktu), bukan dari hubungan fisik langsung. Penurunan OR terhadap waktu ini merupakan perilaku *decline* yang lazim dimodelkan dengan kurva Arps [19] dan mencerminkan faktor yang tidak tercakup oleh ketiga variabel permukaan, seperti perubahan tekanan reservoir, *water cut*, dan *gas-oil ratio*. Selain itu, penelitian ini hanya menggunakan data satu sumur selama 207 hari sehingga hasilnya belum tentu dapat digeneralisasikan pada sumur lain dengan karakteristik reservoir yang berbeda.

4. SIMPULAN

Pada data produksi sumur 15/9-F-12H (207 hari pada tahun 2016), CS merupakan variabel dengan korelasi terkuat terhadap OR ($r = 0{,}58$), diikuti WHT ($r = 0{,}32$) dan WHP ($r = -0{,}32$), dengan multikolinearitas kuat antara WHP dan WHT ($r = -0{,}92$; VIF = 7,16 dan 6,71) yang perlu diperhatikan dalam interpretasi koefisien regresi linier. Pada pembagian acak 80:20, GAM menghasilkan prediksi OR yang lebih baik daripada regresi linier berganda (R^2 uji 0,543 terhadap 0,267; RMSE 32,92 terhadap 41,70 Sm^3/hari; MAE 24,26 terhadap 36,28 Sm^3/hari), yang mengindikasikan hubungan yang tidak sepenuhnya linier. Namun, keunggulan tersebut tidak bertahan pada pembagian berbasis waktu dan *blocked cross-validation*, ketika kedua model menghasilkan R^2 uji negatif karena OR menurun kuat seiring waktu ($r = -0{,}95$) dan perubahan bukaan *choke* pada periode akhir menempatkan data uji di luar rentang data latih. Selain itu, 85,5% data berada pada *choke* terbuka penuh sehingga variasi OR pada kondisi tersebut tidak dapat dijelaskan oleh ketiga variabel permukaan. Oleh karena itu, hasil penelitian ini lebih tepat dipandang sebagai perbandingan interpretasi hubungan di dalam satu periode, bukan sebagai model prediksi yang dapat digeneralisasikan lintas waktu.

Penelitian selanjutnya disarankan memperluas data ke beberapa sumur dan periode yang lebih panjang, menambahkan variabel reservoir dan waktu seperti *water cut*, *gas-oil ratio*, dan hari produksi, menerapkan validasi berbasis waktu pada setiap sumur, serta membandingkan dengan metode lain seperti *random forest* atau *gradient boosting*.

5. DAFTAR PUSTAKA

[1] W. E. Gilbert, "Flowing and gas-lift well performance," in Drilling and Production Practice, New York, NY, USA: American Petroleum Institute, 1954, pp. 126–157.

[2] A. Choubineh, H. Ghorbani, D. A. Wood, S. R. Moosavi, E. Khalafi, and E. Sadatshojaei, "Improved predictions of wellhead choke liquid critical-flow rates: modelling based on hybrid neural network training learning based optimization," Fuel, vol. 207, pp. 547–560, 2017, doi: 10.1016/j.fuel.2017.06.131.

[3] T. A. Ganat and M. Hrairi, "A new choke correlation to predict flow rate of artificially flowing wells," J. Pet. Sci. Eng., vol. 171, pp. 1378–1389, 2018, doi: 10.1016/j.petrol.2018.08.004.

[4] M.-S. Dabiri, F. Hadavimoghaddam, S. Ashoorian, M. Schaffie, and A. Hemmati-Sarapardeh, "Modeling liquid rate through wellhead chokes using machine learning techniques," Sci. Rep., vol. 14, art. 6945, Mar. 2024, doi: 10.1038/s41598-024-54010-2.

[5] R. Alharbi, N. Alageel, M. Alsayil, R. Alharbi, and A. Alhakamy, "Prediction of oil production through linear regression model and big data tools," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 12, pp. 380–387, 2022, doi: 10.14569/IJACSA.2022.0131246.

[6] A. Davtyan, A. Rodin, I. Muchnik, and A. Romashkin, "Oil production forecast models based on sliding window regression," J. Pet. Sci. Eng., vol. 195, p. 107916, 2020, doi: 10.1016/j.petrol.2020.107916.

[7] W. Liu, W. D. Liu, and J. Gu, "Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network," J. Pet. Sci. Eng., vol. 189, p. 107013, 2020, doi: 10.1016/j.petrol.2020.107013.

[8] I. Makhotin, D. Koroteev, and E. Burnaev, "Gradient boosting to boost the efficiency of hydraulic fracturing," J. Pet. Explor. Prod. Technol., vol. 9, no. 3, pp. 1919–1925, 2019, doi: 10.1007/s13202-019-0636-7.

[9] T. Hastie and R. Tibshirani, Generalized Additive Models. London, U.K.: Chapman & Hall, 1990.

[10] M. E. Wigwe, E. S. Bougre, M. C. Watson, and A. Giussani, "Comparative evaluation of multi-basin production performance and application of spatio-temporal models for unconventional oil and gas production prediction," J. Pet. Explor. Prod. Technol., vol. 10, no. 8, pp. 3091–3110, Jul. 2020, doi: 10.1007/s13202-020-00960-y.

[11] S. N. Wood, "Generalized additive models," Annu. Rev. Stat. Appl., vol. 12, pp. 497–526, 2025, doi: 10.1146/annurev-statistics-112723-034249.

[12] J. H. Kim, "Multicollinearity and misleading statistical results," Korean J. Anesthesiol., vol. 72, no. 6, pp. 558–569, Dec. 2019, doi: 10.4097/kja.19087.

[13] Equinor, "Volve field data set," Equinor, Stavanger, Norwegia, 2018. [Online]. Available: https://www.equinor.com/energy/volve-data-sharing. Diakses pada 20 Agustus 2026. Berkas data: Volve_production_2016_per_well.xlsx (lembar NO 15_9-F-12 H); lisensi dan versi data: Equinor Open Data Licence, dirilis Juni 2018.

[14] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Comput. Sci., vol. 7, p. e623, 2021, doi: 10.7717/peerj-cs.623.

[15] D. Servén and C. Brummitt, "pyGAM: Generalised additive models in Python," Zenodo, 2018, doi: 10.5281/zenodo.1208723.

[16] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in Proc. 9th Python in Science Conf. (SciPy), Austin, TX, USA, 2010, pp. 92–96, doi: 10.25080/Majora-92bf1922-011.

[17] C. Cao, P. Jia, L. Cheng, Q. Jin, and S. Qi, "A review on application of data-driven models in hydrocarbon production forecast," J. Pet. Sci. Eng., vol. 212, p. 110296, 2022, doi: 10.1016/j.petrol.2022.110296.

[18] Z. Tariq, M. S. Aljawad, A. Hasan, M. Murtaza, E. Mohammed, A. El-Husseiny, S. A. Alarifi, M. Mahmoud, and A. Abdulraheem, "A systematic review of data science and machine learning applications to the oil and gas industry," J. Pet. Explor. Prod. Technol., vol. 11, no. 12, pp. 4339–4374, 2021, doi: 10.1007/s13202-021-01302-2.

[19] X. Ma and Z. Liu, "Predicting oil production using the novel multivariate nonlinear model based on Arps decline model and kernel method," Neural Comput. Appl., vol. 29, no. 2, pp. 579–591, 2018, doi: 10.1007/s00521-016-2721-x.

Daftar Simbol

RMSE = *Root Mean Square Error*

MAE = *Mean Absolute Error*

VIF = *Variance Inflation Factor*

WHP = Tekanan kepala sumur, bar

WHT = Temperatur kepala sumur, °C

CS = *Choke Size* (bukaan *choke*), %

GAM = *Generalized Additive Model*

OR = Laju alir minyak, Sm^3/hari

EDF = *Effective Degrees of Freedom* (derajat kebebasan efektif)

λ = parameter penalti (*smoothing parameter*) pada GAM