A multi-agent deep reinforcement learning method for CO₂ flooding rates optimization

Li Rongtao¹, Xinwei Liao¹, Xiaoyan Wang², Yang Zhang², Lingyu Mu³, Peng Dong¹ and Kang Tang¹

Abstract
The CO₂ flooding with superior displacement efficiency and high injectivity is an efficient enhanced oil recovery method. However, due to the unfavorable sweep efficiency particularly for strong heterogeneous reservoirs and immiscible flooding, the oil recovery on site is not all favorable. Multi-well rates optimization is one of common measures improving sweep efficiency with easy implement and low cost. There are many rates optimization methods have been proposed by now. In this research, we first introduced the multi-agent deep deterministic policy gradient (MADDPG) algorithm to the multi-well rates optimization of CO₂ flooding, and the new rates optimization method was built. The MADDPG adopts the centralized training and decentralized execution algorithm framework, and overcomes the defect that the single-agent reinforcement learning cannot deal the multi-well rates optimization well and also avoids the dimensional disaster problems. We treated each well as an agent, and each agent has its own reward, state and action. We chose the net present value (NPV) as the reward, the injection-production rate change range as the action element, and the production time, the bottom hole pressure, the oil production rate, and the gas-oil ratio as the state elements. The simulation results show that the optimal case obviously improves the NPV compared with the base case, and the simulation case with strong heterogeneity and immiscible flooding can also converge to the optimal target, which prove the effectiveness and robustness of the rates optimization method respectively. This research provides recommendations that improving the oil recovery by increasing the sweep efficiency to increase the income, and reducing the invalid CO₂ injection to

¹China University of Petroleum (Beijing), School of Petroleum Engineering, Beijing, China
²Oil Production Technology Research Institute, PetroChina Dagang Oil-field, Tianjin, China
³CNPC Engineering Technology R&D Company Limited, Beijing, China

Corresponding author:
Xinwei Liao, China University of Petroleum (Beijing), School of Petroleum Engineering, No.18 Fuxue Road, Changping District, Beijing, 102249, China.
Email: xinwei@cup.edu.cn

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
decrease the cost can achieve the optimal NPV. Reservoir heterogeneity seriously impairs the rates optimization performance, and rates optimization makes little difference to extreme strong interlayer heterogeneity for serious interlayer influence.

Keywords
Rates optimization, multi-agent deep reinforcement learning, sweep efficiency, injection-production connectivity, invalid injection

Introduction
CO₂ flooding is an efficient enhanced oil recovery method, for its high injectivity and superior displacement efficiency. It has been acknowledged that the displacement efficiency of the completely miscible CO₂ flooding is reach to 100% (Holm, 1987). However, the oil recovery of the CO₂ flooding on site is not all favorable, for the inferior displacement efficiency and sweep efficiency (Zhang et al., 2018; Wang et al., 2021). Because the negative effects result from the miscibility degree, the strong heterogeneous reservoir, the adverse mobility ratio, and the gravity override, etc. Many methods were proposed to improve the sweep efficiency, such as profile control, water plugging, injection-production well conversion, and optimal perforation, etc. (Merchant, 2010). The injection and production rates optimization is one of common methods to improve the sweep efficiency, and it has the advantages of low cost, easy implement and favorable effect.

Many researchers have recognized that rates optimization is very necessary for the water flooding, for it can obviously enhance the water flooding oil recovery by improving the sweep efficiency (Van Essen et al., 2009; Udy et al., 2017). Compared with the water flooding, the CO₂ flooding is more needed the rates optimization for its inferior sweep efficiency. The unfavorable sweep efficiency of the CO₂ flooding results from the adverse mobility ratio, because the viscosity of the injected CO₂ is much lower than the injected water. Many researchers have proved that rates optimization can obviously enhance the CO₂ flooding oil recovery, but they didn’t thoroughly investigate the mechanisms behind this phenomenon. Chen and Albert compared the performance difference between the continuous water flooding, the continuous CO₂ flooding and the water-alternating-gas (WAG). They found that the oil recovery of the continuous CO₂ flooding is lower than the WAG for the inferior sweep efficiency, but it’s higher than the continuous water flooding for the superior displacement efficiency. Chen et al. concluded that rates optimization enhances the oil recovery obviously by increasing the sweep efficiency, but they did not give the concrete evidence to prove their opinion. In this paper, we quantitatively characterized the sweep efficiency and further analyzed the mechanisms from different aspects. Other researchers also investigated the effect of the reservoir heterogeneity on the robustness of the proposed rates optimization method (Chen and Reynolds., 2018). However, they rarely considered the influence of the reservoir heterogeneity on the rates optimization performance. We extended their research and found that the rates optimization performance is affected by the reservoir heterogeneity seriously particularly for extreme strong interlayer heterogeneous reservoirs.

There are many traditional injection-production rates optimization methods were proposed, such as genetic algorithm, particle swarm optimization algorithm, conjugate gradient algorithm, and gradient descent algorithm, etc. They are mainly classified into two groups, gradient algorithm and gradient-free algorithm, according to the need of the gradient information of the objective function
for maximum value calculation or not. The advantages of gradient methods, such as conjugate gradient algorithm and gradient descent algorithm, are the high efficiency and low calculation cost (Perry, 1978; Jansen, 2011). However, they have a great disadvantage for getting the gradient information of the objective function that needs the simulator resource code, which is difficult for application and restricts their implication range (Suwartadi, 2012). The gradient-free methods, such as genetic algorithm and particle swarm optimization algorithm, avoid their drawbacks and have wide application range for they can easily cope with all simulators (Fonseca and Fleming, 1993; Souza et al., 2010). While their optimal results are not determined, and the calculation cost is high for low efficiency (Pham et al., 2011). The reinforcement learning method is a gradient-free optimization method, and can cope with any simulators with no need of the simulator resource codes. What’s more, the deep reinforcement learning has its algorithm advantage. When applying the deep reinforcement learning to deal with the similar optimal problem of the other oil field blocks, the use of transfer learning can significantly improve the training efficiency (Khan et al., 2012).

The reinforcement learning is one of optimization methods, and many researchers have applied reinforcement learning methods to injection-production rates optimization and achieved favorable results recently. The traditional Q-learning off-line policy reinforcement learning was used to search the optimal water flooding rates control policy of one injection well and one production well completed in conceptual reservoir model with three layers (Talavera et al., 2010). However, the traditional Q-learning can only deal with the optimal problem with small action and state spaces. The traditional SARAS on-line policy reinforcement learning was also used to optimize the injection rate policy of the steam-assisted gravity drainage (SAGD) (Guevara et al., 2018). The action spaces consist of three discrete actions: increase, no change and decrease in the injection rate. The state spaces consist of the cumulative oil production, the cumulative water production and the cumulative water injection. Although the physical phenomena of the SAGD are complex, the rate optimal calculation amount is small for there is only one injection well rate was optimized. Many deep reinforcement learnings were applied to optimize the injection rate of water flooding under geological uncertainty with one injection well and one production well. It found that the deep Q-learning network (DQN) can converge to the same high net present value (NPV) as the particle swarm optimization method. The deep deterministic policy gradient (DDPG) converges the fastest among the others deep reinforcement learnings: the DQN, the DDQN, and the dueling DDPG (Ma et al., 2019). The state spaces of the deep reinforcement learning not only consist of some handcrafted features, but also consist of the pixel data such as the saturation and pressure distributions (Miftakhov et al., 2020). It shows that the agent can extract the physical characteristics from the inputted pixel data, but costs much time to converge than the handcrafted features such as the well production rate and bottom hole pressure. However, it only considers one injection well and one production well, which is far from realistic condition with many wells. The proximal policy optimization (PPO) algorithm with the convolutional neural network was adopted to optimize the field development, which includes the determination of well number, well locations, and drilling sequence under different scenarios (Nasir et al., 2020, 2021). The state spaces consist of the pixel data for the convolutional neural network is used. Their work has great advantage over traditional gradient-free optimization algorithms, such as genetic algorithm and particle swarm optimization, for it has strong generalization ability and can be easily generalized to different optimal field development scenarios, such as well type, well pattern and water injection policy determination, etc. The multi-well injection rates optimization of water flooding was investigated by using the single agent reinforcement learning, while they assumed that the bottom hole pressures of production wells are constant and only optimized the injection well rates (Hourfar et al., 2017). However, they didn’t treat multi-well rates optimization well by using the single agent
reinforcement learning, and didn’t eliminate the disturbs results from the other wells. Unfortunately, few researchers have studied the rates optimization based on the deep reinforcement learning thoroughly, and no one has applied the MADDPG to the rates optimization of CO₂ flooding by now according to our knowledge.

In this research, the novel CO₂ flooding multi-well rates optimization method based on the MADDPG was built. The effectiveness and robustness of the proposed rates optimization method were proved by two groups simulation experiments with great differences in main influence factors miscibility degree and reservoir heterogeneity. We quantitatively characterized the sweep efficiency during the optimal period and further investigated the main mechanism of the rates optimization from different aspects. The effects of miscibility degree and reservoir heterogeneity on the rates optimization performance were investigated. The simulation results show that the main mechanism of rates optimization is improving the sweep efficiency instead of the displacement efficiency. The miscibility degree and reservoir heterogeneity have significant influence on the rates optimization performance. The rates optimization makes little difference to extreme strong interlayer heterogeneous reservoirs for serious interlayer influence.

Optimal problem descriptions

The CO₂ flooding improves the oil recovery significantly particularly for low permeability reservoirs for its high displacement efficiency and superior injectivity. The CO₂ reaches the supercritical state with low viscosity at the reservoir condition, and the flow behavior follows the Darcy’s law. The phase behavior of the CO₂ flooding is affected by each component (Fanchi, 1987), and the compositional model is built. Each component has its continuity equation according to the principle of material balance. The fluid system properties such as density, viscosity, volume and surface tension are not only the function of the pressure as the black oil model, but also the function of the component state (Wei et al., 2011). The equation of state (EOS) is proposed to describe each component state for further phase equilibrium and fluid property calculation. The EOS coefficients were determined by matching the experimental data and the simulation data, and they are the basis for further reservoir simulation study. Compared with water flooding, the displacement efficiency of CO₂ flooding is more favorable (Holm and Josendal, 1974). The displacement efficiency of CO₂ flooding is proportional to miscibility degree characterized by the surface tension (Claridge, 1972; Cinar and Orr, 2004). When the surface tension between the CO₂ and oil decreases to zero, the fluid system becomes one phase and achieves complete miscibility, and the displacement efficiency is almost high to 100%. However, the miscibility degree is affected by many factors such as pressure, temperature and oil component, and the complete miscibility is difficult to achieved (Ghedan, 2009). The miscibility degree affects the oil and CO₂ phase flow behavior. The oil and CO₂ relative permeability are not only influenced by the corresponding phase saturations, but also by the miscibility degree (Ahmadloo et al., 2009). Other auxiliary equations were also built for describing the other phenomena and solving equations.

There are many factors influencing the oil recovery of CO₂ flooding by reducing the displacement efficiency or the sweep efficiency. Firstly, the viscous fingering caused by the viscosity difference between the CO₂ and the oil (Kargozarfard et al., 2019). The CO₂ with low viscosity cannot displace the oil with high viscosity completely and uniformly, while forms fingers with different shapes at the displacement front as it moves through the porous medium (Suekane et al., 2017). Secondly, the miscibility degree increases the displacement efficiency significantly by reducing the surface tension, and the displacement efficiency is high to 100% theoretically under complete miscibility (Gardner et al., 1981; Zheng et al., 2021). However, the miscibility degree is influenced
by many factors and is difficult to achieve or maintain complete miscibility. Thirdly, the strong heterogeneous reservoirs with ultra-high permeability layers, advantage channels, natural fractures or hydraulic fractures always result in low sweep efficiency (Sayegh et al., 2009). The injected CO2 is inclined to flow along high permeability channels with low seepage resistance, and too early breaks through the production well leading to serious gas channeling and invalid CO2 injection. However, the sweep efficiency of the low permeability zones with high seepage resistance is low, and too much oil is left without recovery. The strong heterogeneous reservoirs with large difference of seepage resistance result in the low sweep efficiency and the invalid CO2 injection. Fourthly, the gravity override results from the fluid density difference and the gravity effect. The injected CO2 with low density is inclined to flow along the top layers, and the oil with high density is prefers to seepage along the bottom layers, which result in gas override and low vertical sweep efficiency (Tiffin and Kremesec, 1988). Fifthly, human factors such as well type, well pattern, production schedule and perforation degree have significantly influence on oil recovery (Alhuthali et al., 2007). The horizontal well has superior productivity compared with the vertical well, while it surfs more serious gas channeling. According to the well type and the reservoir permeability distribution, the corresponding well pattern is determined to achieve favorable sweep efficiency. The mainstream line area has high displacement pressure gradient, which results in high flow velocity and superior sweep efficiency. However, the area with low displacement pressure gradient suffers the contrary condition (Sudaryanto and Yortsos, 2001). Due to the differences of the reservoir heterogeneity and displacement pressure gradient distribution, it is difficult to achieve favorable sweep efficiency particularly for the CO2 flooding with adverse mobility ratio. Therefore, the injection-production rates optimization is proposed to further improve the sweep efficiency without changing the already existing well type and well pattern. It has the advantage of low cost, easy implement and obvious effect. The rates optimization is important and necessary for CO2 flooding, for CO2 flooding has low sweep efficiency while with high displacement efficiency.

Optimal algorithm theory

The multi-well rates optimization of CO2 flooding is an optimization problem, and it mainly consists of three sections: optimization target, constraint conditions and optimal decision. The multi-well rates optimal target of CO2 flooding is the NPV, that is getting the maximum economic benefit. The constraint conditions include keeping the injection-production ratio around 1.0 to maintain the reservoir average pressure around the initial pressure avoiding the formation damage caused by stress sensitivity. The lower limit of the production well bottom hole pressure is higher than the crude oil bubble point pressure. The upper limit of the injection well bottom hole pressure is lower than the reservoir rock fracking pressure. The optimal decision is the optimal injection-production well rates policy during optimal production time. What’s more, many evaluation indexes were used to describe the rates optimization favorable results and to explain the reasons behind the phenomena.

The reinforcement learning method is one of machine learning methods, and the goal of the reinforcement learning is to find the optimal policy to get the biggest cumulative reward. The main traditional reinforcement learning method includes the Q-learning, temporal difference and Monte Carlo. These traditional reinforcement learning can only solve simple problems for the computing capacity constraints. Complex problems with continuous action and state spaces always suffer dimensional disaster. Thanks to the development of computational hardware graphics processing unit, the computing capacity improves significantly and the deep learning becomes realizable (Zhifei and Joo, 2012). Recently many new reinforcement learning methods were developed
based on the deep learning, such as the DQN, the deep double Q-learning network (DDQN), the deterministic policy gradient (DPG), the deep deterministic policy gradient (DDPG) and the MADDPG. The DQN first combines the traditional Q-learning method with the deep learning, and can solve complex problems with large computation amount (Mnih et al., 2013). The DDQN using double networks further improves the DQN by reducing the value overestimation. The double networks decouple the action selection by using the online network from the value evaluation by the target network, which avoids the value overestimation results from both the action selection and the value evaluation by using the same single online network (Van Hasselt et al., 2016). The DQN and the DDQN can deal with complex problems with continuous state spaces for the value function approximation improves significantly with the help of deep learning. However, they cannot deal with the problem with continuous action spaces well, for they cannot enumerate all possible actions to select the highest reward action limited by computation amount (Sutton et al., 1999). Policy gradient algorithm can deal with problems with continuous action spaces well, for it uses a parametric action probability distribution to present the policy mapping the state to the action. It adjusts the policy parameters to get the higher cumulative reward with high computing efficiency without enumerating all possible actions (Peters and Bagnell, 2010). The policy gradient is divided into two types: the deterministic policy gradient (DPG) and the stochastic policy gradient (SPG). The DPG policy maps the state to the deterministic action, while the SPG maps the state to an action probability with stochastic action. The DPG algorithm has superior computing efficiency compared with the SPG algorithm, for the DPG policy gradient only integrates with action spaces requiring less samples, while the SPG policy gradient integrates both action spaces and state spaces needing more samples. What’s more, the DPG can introduce an off policy learning actor-critic algorithm and adopt a stochastic behavior policy to ensure adequate exploration as the SPG (Silver et al., 2014). The DDPG method combines the DPG with the deep learning, and avoids the DPG cannot deal the large problems with continuous states. Therefore, the DDPG can operate the complex problem with continuous action and state spaces well, and avoids the disadvantages of the DQN and the DPG that can only deal large problems with continuous states and actions spaces respectively (Lillicrap et al., 2015). These above reinforcement learning algorithms can only operate optimal problems with single agent well, and they cannot solve optimal problems with multiple agents for not considering the disturbs between each agent and the non-stationary environment. The MADDPG derives from the DDPG, and adopts the modified actor-critic framework with centralized training and decentralized execution (Lowe et al., 2017). The centralized training considers the disturbs and shares the optimized policies of the other agents during training, which is the main reason why the MADDPG can deal optimal problems with multiple agents well.

In this research, we adopted the MADDPG algorithm as the multi-well rates optimization method. The MADDPG algorithm is a gradient-free optimization method, and can be cope with any simulator without getting the simulator resource code. Therefore, the MADDPG algorithm has wide application scope and easy operation. The MADDPG algorithm is one of advanced reinforcement learning methods, which gets the optimal policy by the continuous interaction between the multiple agents and the environment to obtain the best reward. The reinforce learning is based on the Markov chain, which divides the whole optimization process into many steps. The whole process optimal decision consists of the optimal policy of each step. The optimal step policy is determined by the interaction between the agent and the environment at each step. The agent takes action following the policy according the state spaces, which describe the main environment characteristics. The environment provides the corresponding instant reward for the certain action, and then enters the next state. The interaction continues to circulate until the end of the episode. The reward measures the action evaluation, and the policy maps the action to the state. When the highest cumulative
reward is achieved, the optimal policy is achieved, and the corresponding selected action at each step is the best. The agent takes different actions at the same step during different episodes, and achieves different instant rewards. Many samples consisting of \((S_t, a_t, r_t, S_{t+1})\) are created and stored after many episodes. In the following episodes, the agent becomes more and more smart with the help of stored samples that is the intellectual memory. The cumulative reward gradually converges to the optimal value, and the optimal decision also achieves.

We built the multi-well rates optimization method based on the MADDPG algorithm, for the MADDPG can deal the multi-agent optimal problems well. The state spaces and the action spaces of the multi-well rates optimization environment are continuous instead of discrete. The situation number and calculation amount exponentially increase with the increased optimal time step number and the well number. The calculation amount of the rates optimization with many state elements and many time steps particularly for multiple wells is very huge, and always suffers the dimensional disaster. What’s more, the multi-well environment disturbed by the other agents is dynamic and the rates optimal policy is not deterministic, which beyond the application range of the DDPG algorithm only applied to single agent optimal problems. Thanks to the calculation mechanisms of the MADDPG algorithm, each agent of the MADDPG has its own actor network and critic network, and takes action according to its own state but shares the other agents’ policies by the mutual connection of the critic networks as Figure 1 shows. The critic network of each agent has the information that is the optimized policies mapping relations between the state and the action. Thanks to the share of critic networks of each agent, the certain agent learns the other agents’ policies and shares their experiences, which considers the disturbs caused by the other agents and results in the high computation efficiency and convergence speed. Each agent takes action only according to the state of its own observation, and this is on the assumption that the whole environment state is determined when we know the actions of all the other agents at each optimal step. However, this lies a limitation that the optimal situation of each agent is similar, each agent faces the similar situation and the same target, and each agent can share experiences with the other agents.

The MADDPG derives from the DDPG, and extends the application range from single agent to multi-agent conditions that is the main difference between them. The rates optimization method based on the MADDPG takes each well as an agent, and each agent has its own critic network and actor network as the DDPG method shown in Figure 2.

The algorithm processes of the DDPG are as follows:
- Initialize online actor network and online critic network with weights \(\theta^\mu\) and \(\theta^Q\) respectively, and the target networks copy the initial weights of the corresponding online networks.
- Initialize the experience replay memory \(R\).
- for each episode
  - Initialize a random process \(\Omega\) for the action exploration
  - for \(t = 1\) to \(T\), do

  1. According to the online actor network with the exploration noise, an action \(a_t\) was selected.
     \[ a_t = \mu(S_t|\theta^\mu) + N_t \]
  2. The environment executes the action \(a_t\) and achieved the reward \(r_t\) and the next sate \(S_{t+1}\). Store the transition \((S_t, a_t, r_t, S_{t+1})\) in experience replay memory \(R\).
  3. Randomly sample a mini-batch of the \(M\) transitions \((S_i, a_i, r_i, S_{i+1})\) from experience replay memory \(R\).
  4. The target actor network inputs the \(S_{i+1}\) and outputs the \(a_{i+1}\)
  5. The target critic network inputs the \(S_{i+1}\) and \(a_{i+1}\), and outputs the \(Q'(S_{i+1}, a_{i+1})\)
\[ y_i = r_i + \gamma Q'(S_{i+1}, \mu'(S_{i+1}|\theta^\mu)|\theta^Q) \]

⑥ The online critic network inputs the \( S_i \) and \( a_i \), and outputs the \( Q(S_i, a_i) \). Update the online critic network by minimizing the Loss: 
\[
\text{Loss} = \frac{1}{M} \sum_i (y_i - Q(S_i, a_i|\theta^Q))^2
\]

⑦ Update the online actor network based on the sampled policy gradient
\[
\nabla_{\theta^\mu} J \approx \frac{1}{M} \sum_i \nabla a Q(S, a|\theta^Q)|_{S=S_i, a=\mu(S_i)} \nabla \mu(S) \mu(S|\theta^\mu)|S_i
\]

⑧ Soft update the target networks:
\[
\theta^Q \leftarrow \tau \theta^Q + (1 - \tau) \theta^Q' \quad \theta^\mu \leftarrow \tau \theta^\mu + (1 - \tau) \theta^\mu'
\]

end for time step
end for episode

Figure 1. Multi-well rates optimization method framework.
Considering the multi-well rates optimal problem characteristics, we determined the required elements of the MADDPG optimal algorithm. Getting the maximum economic benefit NPV is the optimal target of the rates optimization. The action spaces of the agent are the change range of the injection-production rates. When the cumulative NPV converges to the optimal value, the corresponding action spaces are the optimal injection-production rates adjustment policy. The environment state spaces of production wells consist of the production time, the production well bottom hole pressure, the oil production rate, and the gas-oil ratio. The environment states of injection wells consist of the production time, the injection well bottom pressure, the average oil production rate of its own group production wells, and the average gas-oil ratio of its own group production wells. Considering the characteristic of the MADDPG algorithm, each agent shares experiences with the other agents, the state spaces elements of production wells and injection wells should keep the same kind. Injection wells only have significant influence on the production wells of the group where it stays, so we took the average oil production rate and the gas-oil ratio of its own group production wells as the information elements of state spaces. We chose the production and injection well data instead of the whole reservoir pressure or saturation distribution diagram data, for the well data describe the environment situation more clearly and can easily be achieved with high accuracy. What’s more, the use of well data with less computation amount converges more efficiently than the whole reservoir pressure or saturation distribution diagram data.

The optimal multi-well rates policy is achieved by the continuous interaction between the multiple agents and the environment. Each well as an agent selects a certain action by the actor network following the action selection policy according to the state spaces. The action policy goal is to get the maximum cumulative reward, and the actor network weights change in this direction. The agents transport the selected rates to the environment at each time step, and the reservoir simulator as the environment runs and gets the well data simulation results, which are used for calculating the instant reward NPV and composing the next state spaces elements. The already calculated data as a sample \((S_t, a_t, r_t, S_{t+1})\) stores in the experience pool at each step. In the following episode calculation, the target critic network finds the best action value in the history experience pool, and the
online critic network changes weights to minimize the difference between the history maximum value and the online value. In this way, the optimal target cumulative NPV gradually converges to the optimal value with the increased episodes.

Case studies

We conducted two groups simulation experiments to prove the validity and robustness of the multiwell rates optimal method based on the MADDPG algorithm. Each group has two cases, and one is the base case with the constant injection-production reservoir rates, while the other case with the optimal dynamic injection-production reservoir rates. The difference of the optimal target NPV between the base case and the optimal case proves the validity of the rates optimization method. The two groups simulation experiments with obvious differences in main influencing factors such as reservoir heterogeneity and miscibility degree were set, which were used for illustrating the robustness of the rates optimization method. If the optimal cases of the two groups simulation experiments with significant difference in main affect factors can converge to the optimal value, the rates optimization method has high stability and wide application range. If not so, the optimization method is only applicable for certain conditions.

The multi-well geological model was used by simulation experiments to check the validity and robustness of the proposed rates optimization method. The multi-well geological model is modified from the egg model that is an open geological model adopted by many researchers (Jansen et al., 2014). The sedimentary face of the egg model is fluvial deposit, which has serious plane heterogeneity and is suitable for checking the validity of rates optimization method. To describe the CO₂ flooding rates optimization characteristics more clearly, we reduced the average permeability of the egg model from high permeability 530mD to low permeability 3.41mD and changed the heterogeneous degree as we needed. We used the permeability variation coefficient to describe the heterogeneous degree. The plane permeability variation coefficients of the group 1 and group 2 models are 0.876 and 1.101 respectively, which far beyond the critical value of strong heterogeneity 0.70. The interlayer permeability variation coefficients of the group1 and group 2 models are 0.147 and 1.030, which belong to weak heterogeneity for lower than the critical value 0.50 and extreme strong heterogeneity for much higher than the critical value 0.70 respectively. The group1 has strong plane heterogeneity and weak interlayer heterogeneity, while the group 2 has extreme strong plane heterogeneity and interlayer heterogeneity. The heterogeneous degree of the group 2 is more adverse than the group 1. The interlayer permeability contrast of the group1 and the group 2 are 1.50 and 31.39 respectively. The average permeability, original oil in place and the well pattern of the group1 and group2 are the same for further convenient comparison. According to the position correlation of the permeability and the porosity, the corresponding porosity was also determined. The concrete permeability of the group 1 and group 2 are shown as Figure 3.

We chose the crude oil of X block of the Dagang oil field as the study oil sample, and the crude oil bubble point pressure is 10.67MPa. The compositional simulation model and equation of state were used to describe the fluid system complex phase behavior during the CO₂ flooding. For efficient phase behavior calculation, we grouped all oil components into seven pseudo-components according to the physical similarity as Table 1 shows. We adopted the PR3 state equation to describe the complex phase behavior and the physical properties changed with the pressure, volume and temperature. The PR3 state equation coefficients were achieved by matching the numerical simulation data and the experimental data. The concrete state equation coefficients were shown in Table 2, which were the used for further simulation calculations.
The relative permeability reflecting the interaction between the reservoir rock and the fluid system was achieved by the core flooding experiment. The displacement efficiency of the CO₂ flooding is high to 100% at the complete miscibility condition, which is obvious higher than the water flooding 50%. This is the main reason why the CO₂ flooding even with high cost is very popular on site. The CO₂-oil relative permeability curves are not only the function of the fluid saturation, but also the function of the miscibility degree. The CO₂-oil relative permeability curves are approach to the straight lines at the miscible condition, which has a great difference with the immiscible condition as Figure 4(b) shows. The miscible and immiscible flooding are classified according to the relative size between the displacement pressure and the minimum miscible pressure (MMP). The miscibility degree is characterized by the displacement efficiency, which is influenced by the displacement pressure obviously. For slim tube experiment with one dimensional displacement space, the sweep efficiency is approach to 100%, and the displacement efficiency is almost equal to the oil recovery. As Figure 5 shows that the displacement efficiency of CO₂ flooding increases significantly with the increased displacement pressure from 25MPa to 35MPa, and increases

**Table 1.** Pseudo-component and molar composition.

| Pseudo-component | CO₂ | C1N2 | C2-C6 | C7-C15 | C16-C26 | C27 + |
|------------------|-----|------|-------|--------|---------|-------|
| Molar composition (%) | 0.07 | 24.89 | 16.11 | 26.92 | 17.59 | 14.42 |

**Table 2.** Equation of state pseudo-coefficients.

| Pseudo-component | Molecular weight (g/mol) | Critical pressure (atm) | Critical temperature (K) | Eccentric factors | Equation coefficients Ωₐ | Equation coefficients Ωₐ₀ |
|------------------|--------------------------|-------------------------|--------------------------|------------------|--------------------------|--------------------------|
| CO₂              | 44                       | 73.87                   | 304.7                    | 0.23             | 0.4572                   | 0.0778                   |
| C1N2             | 16.18                    | 45.9                    | 163.1                    | 0.01             | 0.4834                   | 0.0558                   |
| C2-C6            | 50.69                    | 43.12                   | 388.68                   | 0.17             | 0.4218                   | 0.0888                   |
| C7-C15           | 142.52                   | 19.11                   | 702.12                   | 0.4              | 0.4252                   | 0.0949                   |
| C16-C26          | 282.48                   | 14.34                   | 792.32                   | 0.7              | 0.4572                   | 0.0673                   |
| C27 +            | 602.43                   | 6.12                    | 961.09                   | 1.66             | 0.5063                   | 0.0696                   |

The relative permeability reflecting the interaction between the reservoir rock and the fluid system was achieved by the core flooding experiment. The displacement efficiency of the CO₂ flooding is high to 100% at the complete miscibility condition, which is obvious higher than the water flooding 50%. This is the main reason why the CO₂ flooding even with high cost is very popular on site. The CO₂-oil relative permeability curves are not only the function of the fluid saturation, but also the function of the miscibility degree. The CO₂-oil relative permeability curves are approach to the straight lines at the miscible condition, which has a great difference with the immiscible condition as Figure 4(b) shows. The miscible and immiscible flooding are classified according to the relative size between the displacement pressure and the minimum miscible pressure (MMP). The miscibility degree is characterized by the displacement efficiency, which is influenced by the displacement pressure obviously. For slim tube experiment with one dimensional displacement space, the sweep efficiency is approach to 100%, and the displacement efficiency is almost equal to the oil recovery. As Figure 5 shows that the displacement efficiency of CO₂ flooding increases significantly with the increased displacement pressure from 25MPa to 35MPa, and increases

![Figure 3. Group1 and group2 geological model. (a) Group 1 geological model (b) Group 2 geological model.](image-url)
slightly with the increased pressure higher than the MMP 35.12MPa. We set two group simulation experiments to investigate the effects of the miscibility degree. The group1 belongs to the miscible flooding for the initial reservoir pressure 41MPa is higher than the MMP. While the group 2 belongs to the immiscible flooding, because the initial reservoir pressure 25MPa is lower than the MMP and the CO2 displacement efficiency of the core flooding experiment is only 76.67%.

We applied the MADDPG algorithm to the multi-well rates optimal problem, and the optimal process framework as Figure 1 shows. Considering characteristics of the multi-well rates optimal problem and the MADDPG algorithm, we determined the reward, the action and state spaces elements. We took the NPV as the optimal target, and the instant reward calculation as equation (1) shows. The economic parameters used for the NPV calculation were determined. The oil price is 2818 ¥/t, the CO2 injection price is 400 ¥/t, and the produced CO2 treatment price is 30 ¥/t. The base interest rate is 8%, and the rates optimization time step is 3 months. We took the injection-production reservoir rates change range from $-20\%$ to $+20\%$ as the action space element. The injection-production well data with high accuracy as the state spaces elements is easily collected on site. The state spaces consist of the production time, the production well bottom hole pressure, the oil production rate, and the gas-oil ratio. Each well as an agent has its own actor network and critic network, and there are 12 agents with 4 injection wells and 8 production wells. An agent selects a certain action by the actor network according to its own environment state spaces. The actor network consists of four layers, and the interlayer neurons number is 32. The neurons number of the actor network input layer is equal to the single agent state dimension four, and the neurons number of the actor network output layer

Figure 4. Water-oil and co2-oil relative permeability curves. (a) Water-oil relative permeability curves (b) CO2-oil relative permeability curves.

Figure 5. MMP measured curve.
is equal to the single agent action dimension one. Each agent evaluates the target value of the
considered state and action spaces by using the critic network sharing experiences with the other
agents. The critic network consists of four layers, and the interlayer neurons number is 64.
The input layer neurons number of the critic network is equal to the sum of all agent action
and state dimensions 60. The output layer neurons number of the critic network is one, which
is equal to the dimension of the target value. The optimal rates policy is achieved by the continuous
interaction between the agents and the environment, and the optimal cumulative target gradually converges to the optimal value with the increased episodes.

\[ r_t = P_o Q_o - C_{CO_2-INJ} Q_{CO_2-INJ} - C_{CO_2-PRO} Q_{CO_2-PRO} \]  

(1)

where \(P_o\) is the oil price, ¥/t, \(C_{CO_2-INJ}\) and \(C_{CO_2-PRO}\) are the injected CO2 price and the produced CO2 treatment price, ¥/t. \(Q_o, Q_{CO_2-INJ}\) and \(Q_{CO_2-PRO}\) are the oil production total, CO2 injection total and CO2 production total at each rates optimization time step respectively, t.

The multi-well rates optimal results are shown as Figure 6. The optimal target cumulative NPV gradually converges to the optimal value around the 200th episode, and keeps around the optimal value until the next 1000th episode. The optimal target NPV of the group 1 base case and optimal case are 13.02 \(\times 10^8\) ¥ and 14.29 \(\times 10^8\) ¥. The optimal target NPV of the group 2 base case and optimal case are 6.33 \(\times 10^8\) ¥ and 6.68 \(\times 10^8\) ¥. The optimal NPV value of the group 1 and group 2 optimal cases are 9.75% and 5.53% more higher than base cases respectively, which shows the validity of the MADDPG rates optimization method. What’s more, both the optimal cases of group 1 and group 2 with great difference in main influence factors can converge to the optimal value, which shows the rates optimization method has high robustness and wide application.

Figure 6. Optimal target NPV of group 1 and group 2. (a) Group 1 NPV (b) Group 2 NPV.
Results and discussions

The multi-well optimized production-injection rates were achieved based on the proposed rates optimization method as Figure 7 shows, we then further analyzed the reasons behind this phenomenon. The optimization period for the group 1 and group 2 are 12 years and 8 years respectively. There are many great different differences of the optimized multi-well rates between the group 1 and group 2 as Figure 7 shows. For example, the injection well 2 rate decreases with the increased CO₂ injection total in group 1, while increases in group 2. However, the injection well 3 rate increases with the increase of the CO₂ injection total in group 1, but decreases in group 2. We compared the P95 optimal case that is higher than 95% optimized projects with the base case, and found that the P95 optimal case of the group 1 has the higher oil production total, the less CO₂ production total and the lower CO₂ injection total than the base case as Figure 8(a) shows. That is reason why the P95 optimal case has the better NPV, for the higher income and the lower cost The P95 optimal case of group 2 has the slightly higher oil production total than the base case. The CO₂ production total and the CO₂ injection total of the P95 optimal case are almost equal to the base case for the group 2 as Figure 8(b) shows. Thus the NPV of the group 2 P95 optimal case is slight higher than the base case. The concrete oil production rates, the CO₂ production rates and the CO₂ injection rates are described more clearly as Figure 8 shows.

We found that the P95 optimal case has the higher NPV than the base case, for it has the higher oil production total, and the less CO₂ injection total and the lower CO₂ production total. We further studied why the P95 optimal case has the higher oil recovery by comparing the differences of the sweep efficiency and the displacement efficiency between the P95 optimal case and the base case. As Figure 9 shows the sweep efficiency and displacement efficiency of the group 1 is superior to the group 2. Because the adverse reservoir heterogeneity results in the poor sweep efficiency, and the

Figure 7. Optimal multi-well rates of group 1 and group 2. (a) Group 1 optimal multi-well rates. (b) Group 2 optimal multi-well rates.
CO₂ immiscible displacement leads to inferior displacement efficiency for the group 2. The P95 optimal case of group 1 has obvious higher sweep efficiency than the base case, and the lower displacement efficiency than the base case. The P95 optimal case of group 2 has slight higher sweep efficiency than the base case, and the displacement efficiency is almost equal. Thus we concluded that the optimal case improves the oil recovery by increasing the sweep efficiency instead of increasing the displacement efficiency.

The P95 optimal case increases the oil recovery by improving the sweep efficiency instead of the displacement efficiency to further achieve optimal economic benefit. We then analyzed the means by which the optimal case increases the sweep efficiency. As Figure 10 shows the P95 optimal case of group 1 has the lower gas-oil ratio than the base case during the optimization phase. It can be concluded that the P95 optimal case increases the sweep efficiency by decreasing the gas-oil ratio, that is reducing the negative influence results from the gas channeling. The gas channeling happens when the injected gas flows along the high permeability channel with low resistance and too early arrives the production well, which results in the high gas-oil ratio and the low

Figure 8. Optimal oilfield production data of group 1 and group 2. (a) Group 1 optimal oilfield production data. (b) Group 2 optimal oilfield production data.

Figure 9. Optimal sweep efficiency and displacement efficiency of group 1 and group 20. (a) Group 1 sweep and displacement efficiency (b) Group 2 sweep and displacement efficiency.
sweep efficiency (Wang et al., 2021). The gas channeling is influenced by the reservoir heterogeneous degree and fluid viscosity difference seriously. The more adverse reservoir heterogeneity and fluid viscosity difference result in the more serious gas channeling (Al-Wahaibi, 2010). The P95 optimal case of the group 2 slightly decreases the gas-oil ratio compared with the base case, and the corresponding sweep efficiency difference is little.

We have found that the optimal case improves the sweep efficiency by reducing the negative effects caused by the gas channeling, and we then further studied the concrete reasons with the help of the simulation results. As Figure 11 shows that the gas-oil ratio of the production well 2 of the group 1 P95 optimal case reduces significantly from 7653 sm$^3$/sm$^3$ to 1157 sm$^3$/sm$^3$. The gas-oil ratio of other production wells like PRO5, PRO6 and PRO8 also reduce, while the gas-oil ratio of PRO3 PRO4 and PRO7 increase and the PRO1 almost no change compared with the base case. The gas-oil ratio of PRO2 well of the group1 base case is high to 7653 sm$^3$/sm$^3$, because there is a high permeability channel lies between the injection well 2 and the production 2 as Figure 3 shows, and the INJ2 injection rate of the base case remains constant, which lead to serious gas channeling. While for the P95 optimal case, the INJ2 injection rate decreases with the increased production time, which avoids the invalid CO$_2$ injection and decreases the CO$_2$ injection cost What’s more, the optimal INJ1 injection rate, the PRO2 production rate and the PRO4 production rate all increase compared with the base case as Figure 7(a) shows, which builds the injection-production connectivity between the injection well 1, the production well 2 and production well 4 increasing the drainage area and the sweep efficiency obviously as Figure 12 shows. It can be concluded that the group1 optimal case decreases the gas channeling negative effect by reducing the invalid CO$_2$ injection, and increases sweep efficiency by building the injection-production connectivity.

![Figure 10. Optimal oilfield gas-oil ratio of group 1 and group 2. (a) Group 1 gas-oil ratio (b) Group 2 gas-oil ratio.](image1)

![Figure 11. Multi-well gas-oil ratio of group 1 optimal case and base case.](image2)
The optimal case is not obvious superior to the base case for group 2. The NPV and oil recovery of the group 2 P95 optimal case are slightly higher than the base case, and the oil-gas ratio is trivial lower than the base case. As Figure 13 shows that the gas-oil ratio of the PRO3, PRO4 and PRO5 of the P95 optimal case are even significant higher than the base case. The gas-oil ratio of the PRO2 first increases and then decreases obviously and finally increases significantly. It seems that the group 2 P95 optimal case first increases the CO2 injection rate to increase the sweep efficiency, then decreases the injection rate to reduce the gas channeling negative effect as Figure 8 (b) shows. The gas-oil ratio of the whole oilfield block has little difference and the drainage area is almost no change compared the P95 optimal case and the base case as Figure 14 shows. It shows that injection-production rates optimization makes a little difference for group 2. We then further analyzed the reason. We compared the gas distribution of layer 1, layer 4 and layer 7 between the group 1 and group 2 shown as Figure 15, and found that the group 2 has serious inter-layer influence resulted from extreme strong interlayer heterogeneity. The group1 model has strong plane heterogeneity but weak interlayer heterogeneity, and the rates optimization makes a great difference between the optimal case and the base case. While for group 2 with extreme strong plane heterogeneity and interlayer heterogeneity, the rates optimization has little impact for the serious
**Figure 14.** Remaining oil distribution at the end of optimization stage of group 2 optimal case and base case. (a) Group 2 optimal case (b) Group 2 base case.

**Figure 15.** Specific layers CO$_2$ saturation distribution at the end of optimization stage of group 1 and group 2 optimal cases 1st layer 4th layer 7th layer. (a) Specific layers CO$_2$ saturation distribution of group 1 optimal case 1st layer 4th layer 7th layer (b) Specific layers CO$_2$ saturation distribution of group 2 optimal case.
interlayer influence. When the high permeability layer has already filled with full CO₂ distribution, the low permeability layer has little CO₂ distribution. The following injected CO₂ is inclined to flow along the high permeability layers with low seepage resistance and not increase the sweep efficiency of low permeability layers. The whole well rates optimization becomes useless, and the layered injection-production rates optimization is needed, which beyond this paper study scope and needs further study.

Conclusions and recommendations

In this paper, we built a novel multi-well rates optimization based on the MADDPG algorithm, and achieved some findings on CO₂ rates optimization performance mechanisms and recommended two useful policies for engineers on site.

The novel multi-well rates optimization method based on the MADDPG algorithm is a gradient-free optimization method that can cope with any reservoir simulator without getting the resource code to get the gradient information of the objective function for maximum value calculation. Thus it has the advantage of easy implement and wide application. The MADDPG adopts the modified actor-critic framework with centralized training and decentralized execution, and considers the non-stationary environment and the disturbs between each agent, which overcome the defeats of the single agent reinforcement learning algorithms. The validity and robustness of the proposed novel multi-well rates optimization method were proved by two groups simulation cases with obvious differences in reservoir heterogeneity and miscibility degree.

Our new findings on CO₂ rates optimization mechanisms support and extend already exciting opinions proposed by many researchers (Chen et al., 2010; Chen and Reynolds, 2016). They thought CO₂ rates optimization increases the oil recovery mainly by improving the sweep efficiency, but didn’t give the quantitative evidences. We extended their research and quantitatively characterized the sweep efficiency and displacement efficiency during the optimal period. We found that improving the sweep efficiency by building the injection-production connectivity is the dominant mechanism for CO₂ flooding rates optimization to enhance oil recovery, instead of the displacement efficiency which even decreases in the optimal miscible simulation case.

We also found that miscibility degree and reservoir heterogeneity not only significantly affect CO₂ flooding oil recovery, but also obviously impact the CO₂ rates optimization performance. The simulation case with the immiscible flooding and strong interlayer heterogeneous reservoir has the inferior CO₂ rates optimization performance than the simulation case with the miscible flooding and weak interlayer heterogeneity. The CO₂ rates optimization makes little difference to simulation case with extreme strong interlayer heterogeneity for serious interlayer influence.

Based on our findings, we have two important recommendations to share. Firstly, improving the oil recovery to increase the income, and reducing the invalid CO₂ injection to decrease the cost can achieve the optimal NPV. The increased oil recovery results from the improved sweep efficiency that is achieved by building the injection-production connectivity. Secondly, reservoir heterogeneity seriously impairs the rates optimization performance, and the layered injection-production rates optimization is needed for the extreme strong interlayer heterogeneous reservoirs, because the whole well rates optimization makes little difference for serious interlayer influence.

In the future work, we will apply the novel multi-well rates optimization method to solve water-alternating-gas (WAG) problem, which has more favorable performance and wide application. The WAG rates optimization has more optimal parameters that challenges the optimal algorithm ability. Besides, we will use proxy model to significantly decrease the calculation time during the optimal period, for the reservoir simulator costs too much time with thousand episodes of calculation.
Acknowledgements
We would like to thank the help of Science Foundation of China University of Petroleum Beijing, and the editor and anonymous reviewers for their insightful and constructive comments.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Natural Science Foundation of China, “Key scientific problems of seepage law and efficient development of ultra-low permeability reservoirs” (U1762210), and National Major Science and Technology Project, “New technologies for efficient recovery of low permeability and tight reservoirs” (2017ZX05009004).

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science and Technology Major Projects, National Natural Science Foundation of China, (grant number 2017ZX05009004-005, U1762210).

ORCID iD
Li Rongtao https://orcid.org/0000-0001-8277-3303

References
Ahmadloo F, Asghari K and Jamaloei BY (2009) Experimental and theoretical studies of three phase relative permeability. In SPE Annual Technical Conference and Exhibition. OnePetro.
Alhuthali AH, Oyerinde A and Datta-Gupta A (2007) Optimal waterflood management using rate control. SPE Reservoir Evaluation and Engineering 10(05): 539–551.
Al-Wahaibi YM (2010) First-contact-miscible and multicontact-miscible gas injection within a channeling heterogeneity system. Energy and Fuels 24: 1813–1821.
Chen B and Reynolds AC (2016) Ensemble-based optimization of the water-alternating-gas-injection process. Spe Journal 21(03): 786–798.
Chen B and Reynolds AC (2018) CO2 water-alternating-gas injection for enhanced oil recovery: Optimal well controls and half-cycle lengths. Computers and Chemical Engineering 113: 44–56.
Chen S, Li H and Yang D (2010) Optimization of production performance in a CO2 flooding reservoir under uncertainty. Journal of Canadian Petroleum Technology 49(02): 71–78.
Claridge EL (1972) Prediction of recovery in unstable miscible flooding. Society of Petroleum Engineers Journal 12(02): 143–155.
Cinar Y and Orr FM (2004) Measurement of three-phase relative permeability with IFT variation. In SPE/DOE Symposium on Improved Oil Recovery. OnePetro.
Fonseca CM and Fleming PJ (1993) Genetic algorithms for multiobjective optimization: FormulationDiscussion and generalization. In Proceedings of the 5th International Conference on Genetic Algorithms, 93: 416–423.
Fanchi JR (1987) Effect of complex fluid phase behavior on CO2 flood simulation. In SPE Annual Technical Conference and Exhibition. Society of Petroleum Engineers (SPE), 26: 121–131.
Guevara JL, Patel et al. (2018) Optimization of steam injection for heavy oil reservoirs using reinforcement learning. In SPE International Heavy Oil Conference and Exhibition. OnePetro.
Gardner JW, Orr FM and Patel PD (1981) The effect of phase behavior on CO2-flood displacement efficiency. Journal of Petroleum Technology 33(11): 2067–2081.
Ghedan S (2009) Global laboratory experience of CO₂-EOR flooding. In SPE/EAGE reservoir characterization and simulation conference. European Association of Geoscientists and Engineers.

Holm WL (1987) Evolution of the carbon dioxide flooding processes. *Journal of Petroleum Technology* 39(11): 1337–1342.

Hourfar F, Bidgoly HJ, Moshiri B, et al. (2017) A reinforcement learning approach for waterflooding optimization in petroleum reservoirs. *Engineering Applications of Artificial Intelligence* 77: 98–116.

Holm LW and Josendal VA (1974) Mechanisms of oil displacement by carbon dioxide. *Journal of Petroleum Technology* 26(12): 1427–1438.

Jansen JD (2011) Adjoint-based optimization of multi-phase flow through porous media – a review. *Computers and Fluids* 46(1): 40–51.

Jansen JD, Fonseca RM, Kahrobaei, et al. (2014). The egg model—a geological ensemble for reservoir simulation. *Geoscience Data Journal* 1(2), 192–195.

Khan S G, Herrmann G, Lewis F L, et al. (2012) Reinforcement learning and optimal adaptive control: an overview and implementation examples. *Annual Reviews in Control* 36(1): 42–59.

Kargozarfard Z, Riazi M and Ayatollahi S (2019) Viscous fingering and its effect on areal sweep efficiency during waterflooding: An experimental study. *Petroleum Science* 16(1): 105–116.

Lillicrap TP, Hunt JJ, Pritzel A, et al. (2015) Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Lowe R, Wu YI, Tamar A, et al. (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv e-prints*, arXiv: 1706.02275.

Merchant DH (2010) Life beyond 80: A look at conventional WAG recovery beyond 80% HCPV injection in CO₂ tertiary floods. In SPE International Conference on CO₂ Capture, Storage, and Utilization. OnePetro.

Mnih V, Kavukcuoglu K, Silver D, et al. (2013) Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Ma H, Yu G, She Y, et al. (2019) Waterflooding optimization under geological uncertainties by using deep reinforcement learning algorithms. In SPE Annual Technical Conference and Exhibition. OnePetro.

Miftakhov R, Al-Qasim A and Efremov I (2020) Deep reinforcement learning: Reservoir optimization from pixels. In International Petroleum Technology Conference. OnePetro.

Nasir Y (2020) Deep reinforcement learning for field development optimization. *arXiv preprint arXiv:2008.12627*.

Perry A (1978) A modified conjugate gradient algorithm. *Operations Research* 26(6): 1073–1078.

Pham N, Malinowski A and Bartczak T (2011) Comparative study of derivative free optimization algorithms. *IEEE Transactions on Industrial Informatics* 7(4): 592–600.

Peters J and Bagnell JA (2010) Policy gradient methods. *Scholarpedia* 5(11): 3698.

Suekane T, Ono J, Hyodo A, et al. (2017) Three-dimensional viscous fingering of miscible fluids in porous media. *Physical Review Fluids* 2(10): 103902.

Sayegh SG and Fisher DB (2009) Enhanced oil recovery by CO₂ flooding in homogeneous and heterogeneous 2D micromodels. *Journal of Canadian Petroleum Technology* 48(08): 30–36.

Sudaryanto B and Yortsos Y C (2001) Optimization of displacements in porous media using rate control. In SPE Annual Technical Conference and Exhibition. OnePetro.
Talavera A L, Tupac Y J and Vellasco M M (2010) Controlling oil production in smart wells by MPC strategy with reinforcement learning. Society of Petroleum Engineers (SPE) 2: 1408–1419.

Tiffin DL and Kremesec VJ (1988) Mechanistic study of gravity-assisted CO₂ flooding. SPE reservoir Engineering 3(02): 524–532.

Udy J, Hansen B, Maddux S, et al. (2017) Review of field development optimization of waterflooding, EOR, and well placement focusing on history matching and optimization algorithms. Processes 5(3): 34.

Van Essen G, Zandvliet M, Van den Hof P, et al. (2009) Robust waterflooding optimization of multiple geological scenarios. Spe Journal 14(01): 202–210.

Van Hasselt H, Guez A and Silver D (2016) Deep reinforcement learning with double q-learning. Proceedings of the AAAI Conference on Artificial Intelligence 30(01): 26–31.

Wang Z, Zeng L, Yu J, et al. (2021) Key geological factors controlling oil displacement efficiency of CO₂ injection in low-permeability reservoirs. Energy Exploration and Exploitation 39(3): 993–1009.

Wei Y, Chen Z, Satyro M, et al. (2011) Compositional simulation using the advanced peng-Robinson equation of state. In SPE Reservoir Simulation Symposium. OnePetro.

Zhang N, Wei M and Bai B (2018) Comprehensive review of worldwide CO₂ immiscible flooding. In SPE Improved Oil Recovery Conference. OnePetro.

Zhifei S and Joo E M (2012) A review of inverse reinforcement learning theory and recent advances. In 2012 IEEE congress on evolutionary computation. IEEE, 1–8.

Zheng R, Fan Z, Li X, et al. (2021) Phase boundary of CH₄, CO₂, and binary CH₄-CO₂ hydrates formed in NaCl solutions. The Journal of Chemical Thermodynamics, 154: 106333.