Biophysical Journal
Article

Biophysical Society

A data-assimilation approach to predict population dynamics during epithelial-mesenchymal transition

Mario J. Mendez,[1,2] Matthew J. Hoffman,[3] Elizabeth M. Cherry,[3,4] Christopher A. Lemmon,[2] and Seth H. Weinberg[1,2,5,*]

[1]Department of Biomedical Engineering, The Ohio State University, Columbus, Ohio; [2]Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, Virginia; [3]School of Mathematical Sciences, Rochester Institute of Technology, Rochester, New York; [4]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia; and [5]The Dorothy M. Davis Heart and Lung Research Institute, The Ohio State University Wexner Medical Center, Columbus, Ohio

ABSTRACT Epithelial-mesenchymal transition (EMT) is a biological process that plays a central role in embryonic development, tissue regeneration, and cancer metastasis. Transforming growth factor-β (TGFβ) is a potent inducer of this cellular transition, comprising transitions from an epithelial state to partial or hybrid EMT state(s), to a mesenchymal state. Recent experimental studies have shown that, within a population of epithelial cells, heterogeneous phenotypical profiles arise in response to different time- and TGFβ dose-dependent stimuli. This offers a challenge for computational models, as most model parameters are generally obtained to represent typical cell responses, not necessarily specific responses nor to capture population variability. In this study, we applied a data-assimilation approach that combines limited noisy observations with predictions from a computational model, paired with parameter estimation. Synthetic experiments mimic the biological heterogeneity in cell states that is observed in epithelial cell populations by generating a large population of model parameter sets. Analysis of the parameters for virtual epithelial cells with biologically significant characteristics (e.g., EMT prone or resistant) illustrates that these sub-populations have identifiable critical model parameters. We perform a series of in silico experiments in which a forecasting system reconstructs the EMT dynamics of each virtual cell within a heterogeneous population exposed to time-dependent exogenous TGFβ dose and either an EMT-suppressing or EMT-promoting perturbation. We find that estimating population-specific critical parameters significantly improved the prediction accuracy of cell responses. Thus, with appropriate protocol design, we demonstrate that a data-assimilation approach successfully reconstructs and predicts the dynamics of a heterogeneous virtual epithelial cell population in the presence of physiological model error and parameter uncertainty.

SIGNIFICANCE Epithelial-mesenchymal transition (EMT) is a biological process that represents the transdifferentiation of an epithelial cell to a mesenchymal cell, which includes losing epithelial-type cell-cell adhesion and gaining mesenchymal-type enhanced cell motility. Recent experimental studies have shown heterogeneous phenotypical profiles were associated with different responses to EMT-suppressing/promoting parametric perturbations. Data-assimilation is a technique in which observations are iteratively combined with predictions from a dynamical model to provide an improved estimation of system states. We use this approach to improve the accuracy of predictions of cell population responses to time-dependent EMT perturbations. We show that data-assimilation can reconstruct population-specific responses to perturbations in the presence of physiological levels of parameter uncertainty.

INTRODUCTION

Epithelial-mesenchymal transition (EMT) is a fundamental biological process during which an epithelial cell transdifferentiates into a mesenchymal-like cell, losing epithelial cell state characteristics, such as tight cell-cell adhesion, and acquiring mesenchymal cell state characteristics, such as increasing cell motility and enhanced migratory behavior (1,2). The regulation of EMT and its reverse process, mesenchymal-epithelial transition (MET), play a central role in many physiological processes, such as tissue morphogenesis during development, embryogenesis and gastrulation, and wound healing (2–4). On the other hand, the dysregulation of EMT has been observed to be a pivotal factor in pathological conditions, such as cancer metastasis and fibrotic diseases of the liver, kidney, and heart (2). EMT can be induced by

Submitted March 7, 2022, and accepted for publication July 8, 2022.

*Correspondence: weinberg.147@osu.edu

Editor: Jianhua Xing.

https://doi.org/10.1016/j.bpj.2022.07.014

multiple signaling factors, and one of the most potent inducers of this cellular transition is transforming growth factor-β (TGFβ) (5–8). In recent years, studies have highlighted that the presence of multiple intermediate or partial phenotypical cell states (i.e., hybrid E/M states), for which the cell expresses features of both epithelial and mesenchymal cells, is an important characteristic of EMT and MET. Recent work has implicated that these partial EMT states play a critical role in the actualization of collective cell migration in cancer cells (9–11). This is a paradigm shift from viewing TGFβ-induced EMT as an all-or-none switch to a multi-step process, in which the phenotypical cell fates during TGFβ-induced EMT result in an epithelial state (E), an intermediate or partial EMT state or states (P), or a mesenchymal state (M). In the setting of pathological conditions, it is crucial not only to understand what drives EMT to better understand the pathology but to additionally develop techniques to predict the phenotypical fate of cells undergoing EMT, with an eye toward developing effective therapies to alter EMT progression. That is, it would be desirable to develop tools to perturb EMT progression in a predictable manner.

Many signaling pathways that drive EMT in physiological and pathological conditions have been identified (12–14); however, the ability to predict the phenotypical fate of a cell undergoing EMT remains a challenge, in particular at the level of predicting a population of cells. One of the main complications with making such predictions is the limited number of EMT-associated markers that can be observed experimentally in an individual live cell experiment; i.e., measurements that can be made continuously in time without terminating the experiment. More broadly, any subset of experimental measurements inherently provides an incomplete snapshot of the cell state at a given moment in time. Further, recent experimental studies have shown that, within a population of epithelial cells, highly variable phenotypical profiles were associated with different time- and dose-dependent responses to TGFβ during EMT and MET (15–17). This inherent biological variability and limited real-time measurements in in vitro experiments present an obstacle to predict cellular responses during EMT, and, more importantly, how these cellular responses may change in response to specific perturbations to either enhance or suppress EMT.

Computational models of the signaling pathways that regulate EMT are valuable tools for deepening understating of cell signaling mechanisms and predicting cell fates during EMT dynamics. However, computational models inherently face critical obstacles in the process of reconstructing and predicting the responses of a population of cells, as observed in typical in vitro experiments. In particular, simulations tend to be parameterized or fit to successfully reconstruct the "typical" cellular behavior or responses, but fail to capture the variability across a population with heterogeneous responses to particular perturbations. This limitation partially arises due to the fact that model parameter values are often a compilation of means and extrapolations from a wide range of experimental settings and conditions. The presence of this physiological model error can cause long-term computational predictions to greatly deviate from in vitro dynamics, due to even a small degree of uncertainty in parameters and the nonlinear nature of biological systems.

More broadly, the interactions between model simulations and experiments are generally sequential. For example, in many cases, a series of experiments are performed to construct a model and fit parameters. The model is subsequently simulated under specific perturbations, beyond the initial scope of the model development, and experiments are performed to validate or refute the model predictions. In this setting, if the experiments are performed on an inherently heterogeneous population (e.g., many cells), the model is most likely fit based on experimental means or medians. In this work, we perform a series of studies in which experiments and simulations interact in a more direct and iterative manner. Consider the following thought experiment: a heterogeneous cell population undergoes an experimental protocol for a set period of time, during which the model is iteratively "improved" based on the most recent experimental observations (with the definition of improved clarified below). Importantly, this process occurs for each population member individually, such that each member is associated with a different model state and potentially different model parameters. At the end of this set period of time, the response of each population member to a specific perturbation is predicted. In this approach, model and experiment are combined in a way that the response to a perturbation is predicted before it is applied, such that, for example, a subset of the population can be selected based on their predicted response to that perturbation or the magnitude of the perturbation is designed for each individual member to evoke a desired response. That is, the model predictions directly guide experimental design. This study is a proof of concept for this approach that more broadly could have significant applications in fields such as drug design and diagnostics.

The "improvement" noted above is based on a technique known as data assimilation. Data assimilation is an iterative algorithm that uses a Bayesian statistical modeling approach to integrate noisy and sparse experimental observations with high resolution but imperfect dynamical model predictions (18). This statistical tool functions by iteratively updating a previous state estimate of system dynamics (known as the background) based on new observations of the "true" system (albeit noisy measurements) to generate an improved state estimate (known as the analysis), which is the system state maximum likelihood estimate. These updated and improved state estimates are then used as the initial conditions for the dynamical model to generate a forecast to a future time point until the next observation, and the process then repeats iteratively. One critical feature of this data-assimilation approach is that all state variables of the forecasting system are updated and corrected, even if the observations are based on a small

subset of the system; i.e., measurements of the system state are sparse in addition to being noisy. This is feasible, since the dynamics of all state variables are coupled, enabling prediction of the evolution of unmeasured system states.

In our recent work, we conducted a series of in silico experiments that demonstrated that we can incorporate incomplete cell marker measurements (accounting for experimental noise) into a data-assimilation approach to reconstruct TGFβ-induced EMT dynamics of a single virtual epithelial cell in the presence of minimal physiological model error (19). The data-assimilation algorithm accurately predicted cell fates (i.e., phenotypes) and reconstructed both measured and unmeasured key EMT cell marker expression levels. Further, the approach successfully predicted the timing of EMT-associated state transitions. We also identified ideal data-assimilation algorithmic parameters that resulted in the best predictive accuracy. In this current study, we expand on our prior work and perform a series of computational experiments in which we apply a data-assimilation approach to reconstruct EMT dynamics of several distinct heterogeneous virtual cell populations. We mimic the heterogeneity inherently observed in epithelial cell populations by generating a large population of model parameter sets. The cell population is exposed to a time-dependent exogenous TGFβ dose and either an EMT-suppressing or EMT-promoting parametric perturbation. With appropriate protocol design, we demonstrate that a data-assimilation approach can successfully reconstruct and predict the dynamics of a heterogeneous virtual epithelial cell population in the presence of physiological model error and parameter uncertainty. Further, we find that augmenting the data-assimilation approach, specifically incorporating estimation of population-specific critical parameters, improves the prediction accuracy of phenotypical responses to EMT-suppressing or EMT-promoting perturbations.

METHODS

The data-assimilation methodology used in this study comprises four main components: an epithelial virtual cell (the "truth" system), a data-assimilation algorithm (ensemble Kalman filter), a forecasting system, and a set of observations of selected cell markers of the virtual cell. The core regulatory network of the TGFβ-induced EMT dynamics of the virtual cell and the forecasting system are governed by the model proposed by Tian et al. (20) (described below). To establish the performance of the data-assimilation approach, we perform synthetic experiments of the virtual cell response to a time-varying TGFβ dose and signaling perturbation (either EMT-promoting or -suppressing), inducing first EMT and, in some simulation conditions, the reverse EMT process, MET. As described further below, in this synthetic experiment, the true system is described by the same dynamical system model as the forecasting system, such that the predictive accuracy can be quantified. However, parameters between the true and forecasting system greatly differ, representing physiological levels of model error.

During the forecasting process, at the end of a given time interval, the data-assimilation algorithm (described in more detail below) generates the analysis, an improved state estimate of the virtual cell based on 1)a limited and noisy observation of the virtual cell and 2) the predicted virtual cell system state produced by the forecasting system. The improved state estimate then provides the initial conditions for the next forecast. The data-assimilation state estimates are implemented iteratively for each subsequent time interval. A schematic of the data-assimilation process is illustrated in Fig. 1.

Computational model of EMT

We use the mathematical model proposed by Tian et al. to represent the core regulatory network of TGFβ-induced EMT for a single epithelial cell (20). The model consists of a system of nine ordinary differential equations that govern the concentrations of endogenous TGFβ, snail1 mRNA, SNAIL1 protein, miR34, zeb mRNA, ZEB protein, miR200, E-cadherin, and N-cadherin. E-cadherin and N-cadherin are cell markers for epithelial and mesenchymal cell states, respectively. Each differential equation is composed by a basal production term, a degradation rate term, and a translation/production rate governed of a Hill function (Eq. 1). This model describes the TGFβ-induced EMT core regulatory network dynamics as two coupled or cascading bistable switches; i.e., EMT is represented as a two-stage progression. Each bistable switch is regulated by a double-negative feedback loop, and these are governed by the interplay between the production of transcription factors SNAIL 1/2 and ZEB1/2, and their respective inhibitors, the microRNAs miR-34 and miR-200. Model initial conditions (corresponding to the epithelial phenotype) and baseline parameters are given in Tables S1 and S2, respectively.

$$\frac{d[\mathrm{T}]}{dt} = k_{0T} + \frac{k_T}{1 + \left(\frac{[\mathrm{R200}]}{J_T}\right)^{n_{r200}}} - k_{dT}[\mathrm{T}], \qquad (1a)$$

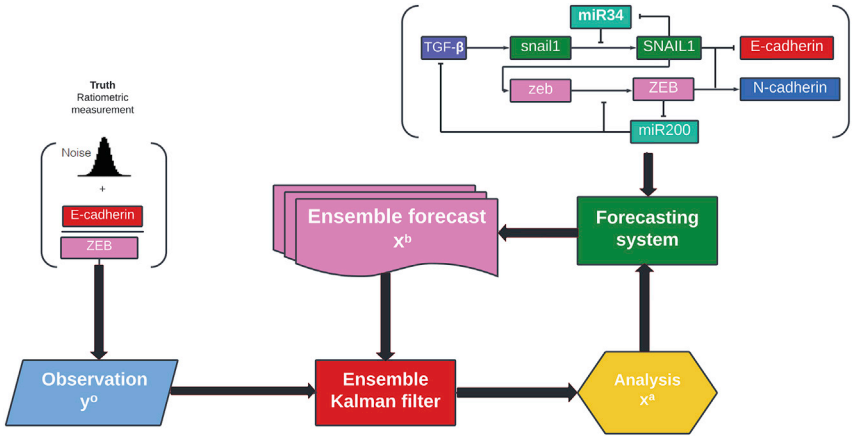

FIGURE 1 Schematic of the data-assimilation approach. Synthetic observations are generated from ratiometric measurements of E-cadherin-to-ZEB from the truth system, plus the addition of Gaussian noise. The numerical model governing EMT dynamics (20) generates forecast ensembles. Combining the model forecasts and noisy observations, the ensemble Kalman filter yields the maximum likelihood estimator for the system state (the analysis), which provides initial conditions for the next iteration. To see this figure in color, go online.

$$\frac{d[s]}{dt} = k_{0s} + k_s \frac{\left(\frac{[T]+[T]_e}{J_s}\right)^{n_t}}{1+\left(\frac{[T]+[T]_e}{J_s}\right)^{n_t}} - k_{ds}[s], \quad (1b)$$

$$\frac{d[S]}{dt} = \frac{k_S[s]}{1+\left(\frac{[R34]}{J_S}\right)^{n_{r34}}} - k_{dS}[S], \quad (1c)$$

$$\frac{d[R34]}{dt} = k_{03} + \frac{k_3}{1+\left(\frac{[S]}{J_{13}}\right)^{n_s}+\left(\frac{[Z]}{J_{23}}\right)^{n_z}} - k_{d3}[R34], \quad (1d)$$

$$\frac{d[z]}{dt} = k_{0z} + k_z \frac{\left(\frac{[S]}{J_z}\right)^{n_s}}{1+\left(\frac{[S]}{J_z}\right)^{n_s}} - k_{dz}[z], \quad (1e)$$

$$\frac{d[Z]}{dt} = \frac{k_Z[z]}{1+\left(\frac{[R200]}{J_Z}\right)^{n_{r200}}} - k_{dZ}[Z], \quad (1f)$$

$$\frac{d[R200]}{dt} = k_{02} + \frac{k_2}{1+\left(\frac{[S]}{J_{12}}\right)^{n_s}+\left(\frac{[Z]}{J_{22}}\right)^{n_z}} - k_{d2}[R200], \quad (1g)$$

$$\frac{d[E]}{dt} = \frac{k_{E1}}{1+\left(\frac{[S]}{J_{E1}}\right)^{n_s}} + \frac{k_{E2}}{1+\left(\frac{[Z]}{J_{E2}}\right)^{n_z}} - k_{dE}[E], \quad (1h)$$

$$\frac{d[N]}{dt} = \frac{k_{N1}}{1+\left(\frac{[S]}{J_{n,1}}\right)^{n_s}} + \frac{k_{N2}}{1+\left(\frac{[Z]}{J_{n,2}}\right)^{n_z}} - k_{dN}[N]. \quad (1i)$$

The core regulatory network of TGFβ-induced EMT is activated by exogenous TGFβ ($[T]_e$), which increases the production of snail1 mRNA ([s]), activating the first double-negative feedback loop and upregulating the translation of SNAIL1 protein ([S]). Increased production SNAIL1 protein inhibits miR-34 ([R34]) production, the inhibitor of SNAIL1 translation, completing the first double-negative feedback loop. SNAIL1 activates the second double-negative feedback loop by increasing the production of zeb mRNA ([z]), upregulating translation of ZEB protein ([Z]), which inhibits the production of the inhibitor of ZEB translation, miR-200 ([R200]). SNAIL1 and ZEB suppress epithelial marker E-cadherin ([E]) and promotes mesenchymal marker N-cadherin ([N]). Finally, the suppression of miR-200 promotes endogenous TGFβ production, enhancing a positive feedback that promotes the irreversibility of the EMT progression for baseline model parameters.

Simulations with baseline model parameters demonstrating the representative EMT transdifferentiation of a virtual cell from an epithelial to a mesenchymal phenotypical state are shown in Fig. 2, for varying exogenous TGFβ doses and epithelial cell initial conditions. For a constant low TGFβ dose (blue), there are minimal changes in SNAIL1, ZEB, E-cadherin, and N-cadherin. For a constant moderate TGFβ dose (red), SNAIL1 is upregulated but with minimal change in ZEB, which yields expression of both E-cadherin and N-cadherin, consistent with a partial or hybrid E/M state. For a constant high TGFβ dose (green), SNAIL1, ZEB, and N-cadherin are maximally enhanced, while E-cadherin is fully suppressed, consistent with a mesenchymal state.

The steady-state N-cadherin levels are shown as a function of exogenous TGFβ dose in Fig. 2 *E* for different cell state initial conditions. Changing TGFβ dose from either epithelial (blue) or partial state (red) initial conditions results in a step-like transition from epithelial-to-partial-to-mesenchymal states, albeit with some initial state dependence (i.e., hysteresis). However, for baseline model parameters, the initial mesenchymal state (green) is irreversible; i.e., N-cadherin levels remain high for all TGFβ doses, including the absence of TGFβ.

Data-assimilation

We utilize a class of algorithms denoted as data-assimilation methods that are used to improve the state estimation and forecasting of dynamical systems by combining observations of the system with numerical model-derived forecasts. Applications of data-assimilation algorithms are well established in the atmospheric science community (21–25) in the development of numerical weather prediction. Several data-assimilation approaches have been developed to estimate and reconstruct selected parameters of a true system, increasing the utility of such algorithms (26–40).

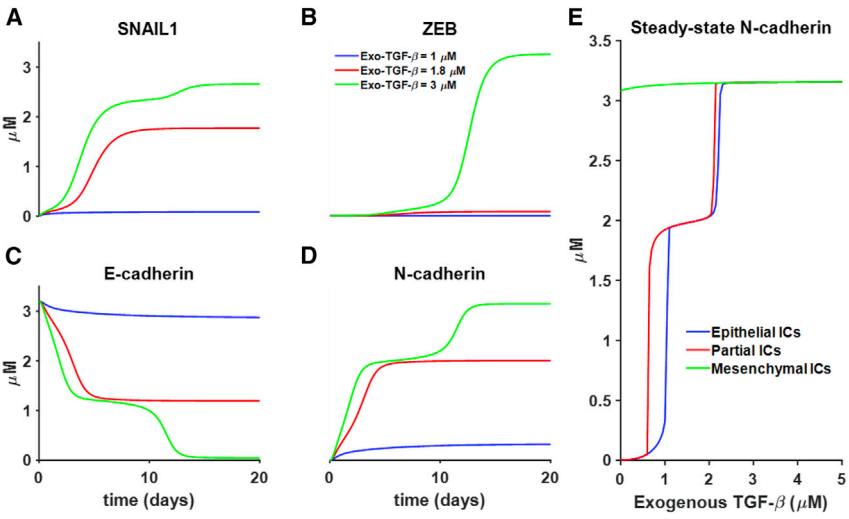

FIGURE 2 TGFβ induces EMT via a partial or intermediate EMT-state transition. (*A–D*) The time course of key epithelial and mesenchymal markers are shown as a function of time, following the addition of 1, 1.8 and 3 μM constant exogenous TGFβ dosages. (*E*) The expression level of N-cadherin on day 20 is shown as a function of the exogenous TGFβ dose, for different initial conditions (ICs). The step-like response illustrates distinct cell states, corresponding with epithelial, partial or intermediate EMT, and mesenchymal states. To see this figure in color, go online.

Ensemble Kalman filter

In this work, we use the ensemble transform Kalman filter (ETKF), which is an extension of the linear Kalman filter for nonlinear problems (18). The ETKF offers an estimate of the most likely state of the system given a prior estimate of the state, a set of observations of the system (potentially sparse and noisy), and uncertainty estimates for both the observations and the state space predictions.

For this problem, the state space at time t is a column vector of the nine model variables at this time:

$$x^t(t) = ([\mathrm{T}](t), [\mathrm{s}](t), [\mathrm{S}](t), [\mathrm{R34}](t), [\mathrm{z}](t), [\mathrm{Z}](t), [\mathrm{R200}](t), [\mathrm{E}](t), [\mathrm{N}](t))^T. \quad (2)$$

In most simulations, we expand this approach to also incorporate parameter estimation; i.e., a specific parameter is also estimated during the data-assimilation process. To do so, the state space vector is augmented to include the specific parameter $p(t)$,

$$x^t(t) = ([\mathrm{T}](t), [\mathrm{s}](t), [\mathrm{S}](t), [\mathrm{R34}](t), [\mathrm{z}](t), [\mathrm{Z}](t), [\mathrm{R200}](t), [\mathrm{E}](t), [\mathrm{N}](t), p(t))^T, \quad (3)$$

where $p(t)$ is the selected parameters to be reconstructed by the data-assimilation algorithm and has trivial dynamics; i.e., $dp/dt = 0$.

For a given time step n, where $t = n\Delta t_{obs}$, where Δt_{obs} is the time interval between observations, the prior state space estimate is produced by an ensemble of forecasting systems called the background state and is denoted x_n^b. That is, multiple versions of the EMT model, or an ensemble, are simultaneously simulated to estimate the background state, and $x_n^{b(i)}$ represents the state space vector of the ith ensemble member. The estimation of the uncertainty on the forecast state space is denoted $\mathbf{P}_n^b$. To generate this estimation of the background uncertainty, it is assumed that the probability distribution of the ensemble of forecasts is Gaussian and the mean and covariance are parameterized by a small number of model states. A similar assumption is made in the Monte Carlo approach; one critical difference is that the ETKF uses fewer ensemble members to fully sample the space.

Following the notation of Hunt et al. (18), given a set of background states $x_n^{b(i)}$, the background is computed as the mean of the ensemble members,

$$x_n^b = \frac{1}{k}\sum_{i=1}^{k} x_n^{b(i)}, \quad (4)$$

where k is the ensemble size, i is the index of the ensemble members, and the covariance is given by the ensemble sample covariance,

$$\boldsymbol{P}_n^b = \frac{1}{k-1}\sum_{i=1}^{k}\left(x_n^{b(i)} - x_n^b\right)\left(x_n^{b(i)} - x_n^b\right)^T. \quad (5)$$

The ETKF determines the state that minimizes the cost function

$$J(\tilde{x}_n) = \left(\tilde{x}_n - x_n^b\right)^T\left(\boldsymbol{P}_n^b\right)^{-1}\left(\tilde{x}_n - x_n^b\right) + \left[y_n^o - H(\tilde{x}_n)\right]^T\boldsymbol{R}_n^{-1}\left[y_n^o - H(\tilde{x}_n)\right], \quad (6)$$

where H is a map from the model space to the observation space (which is typically lower dimensional), y_n^o is the vector of observations, and $\mathbf{R}_n$ is the covariance of these observations. The analysis, denoted as $\boldsymbol{X}_n^a$, represents the state that minimizes the cost function in the subspace spanned by the ensemble members. The analysis error covariance matrix in ensemble space, $\tilde{\mathbf{P}}_n^a$, can be computed in ensemble space as $\tilde{\boldsymbol{P}}_n^a = [\rho^{-1}(k-1)\boldsymbol{I} + \boldsymbol{Y}_n^{bT}\boldsymbol{R}_n^{-1}\boldsymbol{Y}_n^b]^{-1}$, where ρ is a multiplicative inflation parameter that allows the algorithm to compensate for the fact that the small ensemble size tends to lead to underestimation of the true background uncertainty. Multiplying the covariance matrix by a constant greater than 1 (ρ here) is a computationally efficient way of correcting for this underestimation. The inflation factor ρ is a tunable parameter for the assimilation algorithm. The columns of the $\mathbf{Y}_n^b$ matrix are the perturbations of the background ensemble members mapped into observation space. Mathematically, the jth column of $\mathbf{Y}_n^b$ is $y_n^{b,j} = H(x_n^{b(j)}) - y_n^b$, where $y_n^b = \frac{1}{k}\sum_{j=1}^{k} H(x_n^{b(j)})$ is the mean of the background ensemble in observation space. Here, unless otherwise stated, the map H takes the form $H(x_n^{b(j)}) = [\mathrm{E}]_n^{(j)}/[\mathrm{Z}]_n^{(j)}$, where $[\mathrm{E}]_n^{(j)}$ and $[\mathrm{Z}]_n^{(j)}$ are the concentrations of E-cadherin and ZEB from the jth ensemble at time step n (described further below).

The analysis covariance is then used to transform the background ensemble perturbations into analysis ensemble perturbations according to $\boldsymbol{X}_n^a = \boldsymbol{X}_n^b[(k-1)\tilde{\boldsymbol{P}}_n^a]^{1/2}$. Finally, the new analysis mean is computed as

$$\boldsymbol{x}_n^a = \boldsymbol{x}_n^b + \boldsymbol{X}_n^b\tilde{\boldsymbol{P}}_n^a\boldsymbol{Y}_n^{bT}\boldsymbol{R}_n^{-1}\left(\boldsymbol{y}_n^o - \boldsymbol{y}_n^b\right). \quad (7)$$

The analysis mean is added to each column of $\boldsymbol{X}_n^a$ to generate the analysis ensemble members. The analysis ensemble members then become the initial conditions for the next set of forecasts using the EMT model (Eq. 1), simulated for duration Δt_{obs} and producing the next background states $x_{n+1}^{b(i)}$, and then this entire process is iteratively repeated (i.e., incorporating the subsequent observation y_{n+1}^o to calculate x_{n+1}^a). Numerical simulations are performed in MATLAB (Mathworks, Natick, MA) using the ode15s ordinary differential equation solver. Descriptions of the variables in the ETKF method are provided in Table S3 (removing the subscript n for clarity). A more detailed description of the algorithm, including derivations, can be found in Hunt et al. (18). Based on our prior work (19), in this study we utilize ensemble size k of 50, multiplicative inflation factor ρ of 1.4, and observation interval (i.e., intervals between analysis steps) Δt_{obs} of 6 h.

Generation of cell population

We generate a synthetic heterogeneous population of epithelial cells, following our prior approach and others (19,41). Specifically, a large population of model parameter sets is generated to reproduce the biological variability in phenotypical cell states that is observed in in vitro experiments, in which multiple states (i.e., epithelial, partial, or mesenchymal state) are observed in the cell population in response to a specific TGFβ dose concentration and duration (16). To generate a population of model parameter sets, random scaling factors are chosen from a log-normal distribution (with median of 1 and distribution parameter $\sigma = 0.075$) and multiplied with the baseline parameter values for key parameters, specifically basal and regulated production, transcription, translation rates, and degradation rates, for endogenous TGFβ, snail1 mRNA, SNAIL1, miR-34, zeb mRNA, ZEB, and miR-200. We draw 5000 random parameter sets that represent a heterogeneous population of virtual epithelial cells (Fig. 3), illustrated by phenotypical variation for different TGFβ dose and duration. As described in more detail below, subsets of this population are identified, and the accuracy of phenotypical state predictions in response to time-varying TGFβ doses and EMT perturbations are measured.

Numerical experiments

For a given data-assimilation trial, we perform a computational simulation for which we reconstruct the EMT cellular dynamics of a virtual epithelial cell and predict its response to a parametric perturbation. Both the true virtual cell and forecasting system are governed by the core regulatory network model of TGFβ-induced EMT proposed by Tian et al. (20). The

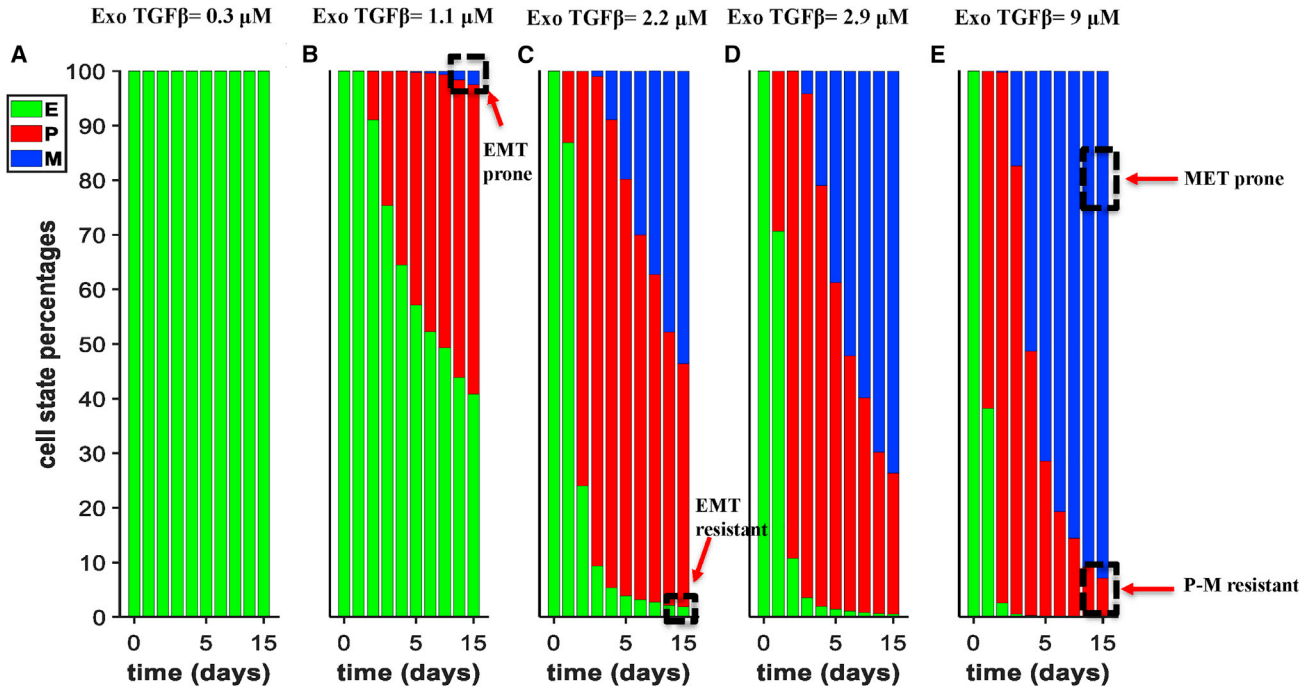

FIGURE 3 Population of model parameter sets reproduces experimental heterogeneity in cell state. (*A–E*) The percentages of cells in the epithelial (E, green), partial (P, red), and mesenchymal (M, blue) state are shown as a function of time for increasing exogenous TGFβ doses (0.3, 1.3, 2.2, 2.9, and 9 μM). We identify four cells sub-populations (dashed line boxes): EMT prone, EMT resistant, P-M resistant, and MET-prone cells. To see this figure in color, go online.

true virtual cell simulations use model parameters that are drawn from the cell population, as described above. However, to mimic parameter uncertainty, the forecasting system ensemble member simulations use baseline model parameters. That is, while the governing dynamics of the true system and forecasting system are the same, nearly all key model parameters differ between the true and forecasting systems. Specifically, we vary 28 out of 45 parameters, with the remaining unchanged 17 parameters comprising Hill coefficients and parameters for E-cadherin and N-cadherin, which can be regarded as system outputs that do not feedback on the overall system dynamics. In each trial, the virtual cell is exposed to a series of time-varying inputs of exogenous TGFβ doses and, at a selected time point, an EMT-suppressing or EMT-promoting perturbation is applied to the virtual cell. We note that the time-varying TGFβ dose was applied to specifically enhance the heterogeneous response of the population subset. The data-assimilation algorithm reconstructs and predicts the EMT dynamics of a given virtual cell at set time intervals. As described above, at the end of each such time intervals, the data-assimilation algorithm generates an improved state estimate of the truth system using an observation of the virtual cell (y_n^o) and the forecast state (x_n^b).

The data-assimilation algorithm iteratively reconstructs both the system state and (in most simulations) a specified parameter. EMT-promoting or EMT-suppressing single parametric perturbations were applied to the virtual cell at a selected time threshold (day 30). The perturbations were applied by multiplying the selected parameter value of the virtual cell by a scaling factor. Importantly, at this same time threshold, system state corrections from the data-assimilation algorithm are no longer applied, such that prediction accuracy of the final cell phenotype is determined from the forecast of the last analysis step. Thus, this experimental design represents observing a cell population subset for a period of time and predicting the response of the population to a specific perturbation only based on observations *before* the perturbation.

The virtual cell was initialized with all state variables in the epithelial state. To initialize each ensemble member of the background, a separate model simulation was performed, with a random TGFβ dose (uniformly sampled between 0 and the given dose for that trial), for a random duration (uniformly sampled between 0 and 20 days). The final state variable concentrations were chosen for the ensemble initial state. Similarly, the estimation of the selected parameter was initialized by multiplying the true parameter value of the virtual cell with a scaling factor for each ensemble member. The value of the scaling factor was sampled from a logarithmic uniform distribution, such that the scaling factor was distributed between 0.5 and 2. To emulate realistic experimental measurements, the observation is defined by the ratio between the concentrations of the epithelial cell marker E-cadherin and the transcription factor ZEB, based on a recent novel stable dual-reporter fluorescent sensor of these two key EMT regulating factors (42). The choice of this observation measurement, i.e., ratiometric measures from a recently developed reporter, is motivated by the possibility of performing data-assimilation studies in real epithelial cells. Observational measurement noise or error was reproduced by adding Gaussian noise of mean 0 and standard deviation 10% of the true ratio magnitude. Minimum E-cadherin and ZEB concentrations were set to 1.1×10^{-50} μM, to avoid negative or undefined ratio values.

To assess the accuracy of the reconstruction of the virtual cell dynamics for a given data-assimilation trial, we calculate the root-mean-square deviation (RMSD) between the true system and the average of the analysis ensembles, summing over all state variables, as a function of time:

$$RMSD(t) = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(x_j^a(t) - x_j^t(t)\right)^2}, \qquad (8)$$

where $x_j^a(t)$ and $x_j^t(t)$ are the jth element of the analysis and truth m-dimensional vectors, respectively, at time t. We calculate the area under the RMSD versus time curve to quantify error for a single trial. Further, the accuracy of the long-term predictions generated by our data-assimilation approach is determined by comparing the predictions of the final phenotypical cell state with the true virtual cell state (i.e., epithelial, partial, and mesenchymal).

RESULTS

Sub-population identification and characteristics

A primary objective of this in silico study is to use a data-assimilation approach to predict and reconstruct cellular systems responses to specific perturbations in the presence of physiological error and population heterogeneity. As noted above, we first generate a population of model parameters to reproduce the phenotypical variability observed experimentally in a population of epithelial cells (16). We simulate the virtual cell population response to a series of constant dosages of exogenous TGFβ for a duration of 15 days, measure the phenotype (epithelial, partial, mesenchymal) based on N-cadherin expression levels, and quantify the changes of the percentages of virtual cells in each cell state (Fig. 3).

For a low dose, all virtual cells remain in the epithelial state (Fig. 3 *A*). For a moderately low dose (Fig. 3 *B*), we observe more heterogeneous phenotypes, with most virtual cells in either the epithelial or partial state by day 15. Importantly, we identify the presence of a small fraction of 125 virtual cells (2.5% of the population) that undergo the full EMT progression, with a final mesenchymal phenotypical state, and categorize this sub-population of virtual cells as EMT prone. For a slightly larger moderate dose (Fig. 3 *C*), we again observe population heterogeneity, with most virtual cells in either the partial or mesenchymal steady state by day 15. However, we also identify a fraction of 95 virtual cells (1.9% of the population) that remain in the epithelial state and categorize this sub-population as EMT-resistant cells. For a large dose (Fig. 3 *D* and *E*), a progressively larger fraction of the population progresses to the mesenchymal state. In spite of a high-dose exogenous TGFβ, we identify a sub-population of 345 cells (6.9% of the population) that remains in the partial state, categorized as P-M resistant. Additionally, for each cell that progresses to the mesenchymal state, we perform an additional simulation in which exogenous TGFβ is removed once the population of cells reaches the mesenchymal state. For the baseline model parameters, as noted above, mesenchymal cells irreversibly remain in this state even after removing TGFβ. However, within this population, a subset of 170 virtual epithelial cells (3.4% of the population) undergo MET, identified as MET-prone cells. Note that there is small overlap between the P-M-resistant cells and the EMT-resistant sub-populations; similarly, we observed a small overlap between MET-prone cells and the EMT-resistant cells. For the remainder of the study, we focus on these four sub-populations, as they each have highly identifiable features of biological interest.

We next compare the parameters associated with the four sub-populations of virtual cells: MET prone, EMT prone, EMT resistant, and P-M resistant. Since all population parameters are scaled versions of the baseline parameters, we analyze the scaling factor distribution for each parameter for each sub-population and statistically compare with the baseline (corresponding to a value of 1) using a series of unpaired *t*-tests. A boxplot for each parameter and sub-population is shown in Fig. 4. For each sub-population, we denote the specific parameters that differ from the baseline as critical parameters (shown as red boxplots).

We find that the number of critical parameters varies with cell sub-population, such that, for EMT-prone cells, 20 out of the 28 parameters are critical, while, for EMT-resistant cell, only eight parameters are critical. Interestingly, we find both similarities and differences in the critical parameters between the sub-populations. The production and degradation rates of the mRNA snail1 and the degradation rate of the protein snail1 are critical parameters for all four sub-populations. All cell sub-populations have at least three critical parameters that are degradation rates, with all degradation rates critical for EMT-prone cells (Fig. 4 *A*). We also find that most of the baseline production rates are non-critical parameters. Interestingly, MET-prone cells are the only sub-population for which endogenous TGFβ production rate (kt) is a critical parameter, with a lower production rate compared with the baseline value (Fig. 4 *D*), consistent with the reversibility of the mesenchymal state. Further, there is quite a bit of variability in the Hill coefficient parameters.

Sub-population heterogeneity

We next investigate the EMT dynamics of the four identified sub-populations in response to a time-varying exogenous TGFβ dose. The time-varying dose is applied primarily for two reasons: 1) the time-dependent changes in TGFβ are comparable with a series of perturbations that produce heterogeneity even within each sub-population (Fig. 5); and 2) the changes in TGFβ induce multiple phenotypic transitions, which enhances the data-assimilation reconstruction. We first consider the MET-prone sub-population of virtual cells, which we expect to be the most difficult to reconstruct, given that the baseline model predicts irreversible EMT dynamics (i.e., the baseline model does not undergo MET).

We apply a time-dependent dose of exogenous TGFβ to each virtual cell in the MET sub-population across a simulation of 40 days (Fig. 5 *A*). An initial moderate TGFβ dose results in some cells transitioning to the partial state, followed by a higher dose, which ultimately results in nearly all cells in the mesenchymal state by day 20, albeit with each virtual cell exhibiting a heterogeneous N-cadherin expression and phenotypic time course (Fig. 5 *B* and *C*). Removal of TGFβ results in MET, again with significant sub-population variability in timing, followed by a moderate dose that produces a mixed sub-population comprising virtual cells in all three states by day 40. We conduct similar time-varying TGFβ doses for the other sub-populations to produce heterogeneous sub-populations (Fig. S1).

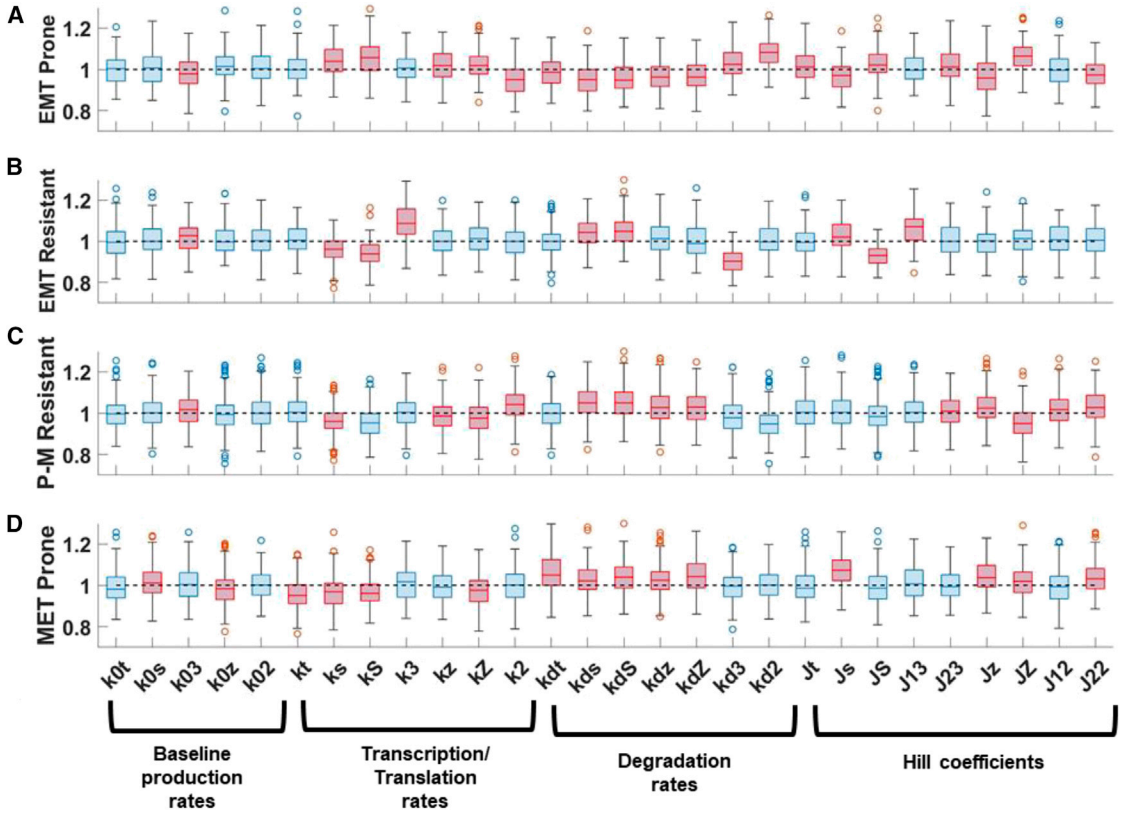

FIGURE 4 Critical parameters deviate from the baseline model for distinct sub-populations. (*A–D*) Boxplots display the parameter scaling factor (relative to baseline of 1, dashed horizontal black line) distribution for four identified sub-populations (EMT prone, EMT resistant, P-M resistant, and MET prone). Critical parameters that significantly deviate from baseline are shown in red, and non-critical parameters are shown in blue. To see this figure in color, go online.

Data-assimilation reconstruction of heterogeneous sub-populations

We next perform data reconstruction during the time-varying TGFβ dose, using a forecasting system with baseline parameters, for each virtual cell of the MET sub-population. One critical aspect of this reconstruction is the definition of a data-assimilation correction window, which specifically defines the time period (days 0–30) for which the data-assimilation algorithm is applied. On day 30, the final data-assimilation reconstruction essentially defines a new set of initial conditions, which are utilized to simulate the final 10 days of the simulations (days 30–40). Note that the end of this window also corresponds with the timing

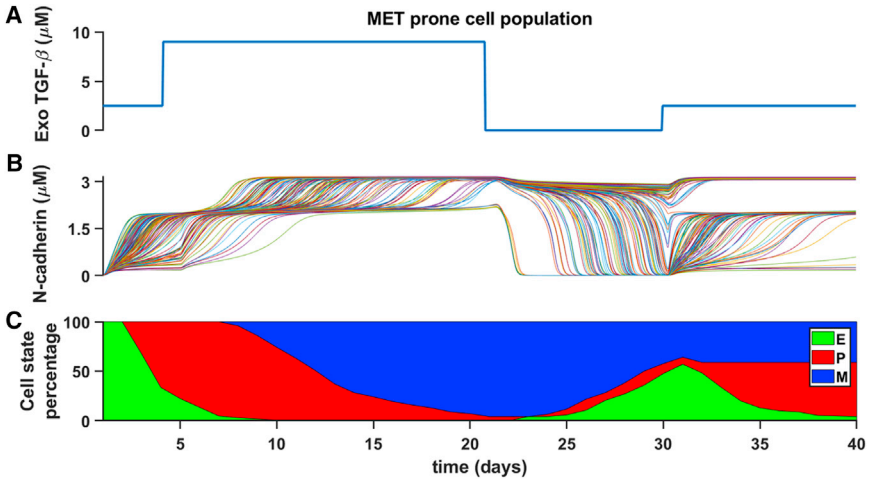

FIGURE 5 MET-prone cells display heterogeneous responses to time-dependent exogenous TGFβ dose. (*A*) Time course of the exogenous TGFβ dose applied to the MET-prone sub-population as function of time. (*B*) N-cadherin expression for the MET-prone cell sub-population members is shown in response to the time-dependent exogenous TGFβ dose. (*C*) The percentage of MET-prone cells in the epithelial (E, green), partial (P, red), and mesenchymal (M, blue) state are shown during the time-dependent exogenous TGFβ protocol. The sub-population exhibits heterogeneous phenotype throughout the in silico experiment. To see this figure in color, go online.

of the TGFβ dose change from none to moderate TGFβ. Thus, this protocol is comparable with observing a cell during a defined correction period, applying a final experimental condition change (here, the return of a moderate TGFβ dose) and predicting the cell response 10 days later. Thus, one could in practice apply this protocol in parallel, in which many cells are simultaneously observed during an initial observation/correction period and then a subset of cells is selected based on their predicted response to additional perturbation(s) at some later time point. Later analysis will consider more complex experimental condition changes, including both changes in TGFβ dose and an EMT-suppressing or -promoting perturbation.

We illustrate three representative virtual cells (cells 15, 2, and 7) from the MET-prone cell sub-population, which reach a final epithelial, partial, and mesenchymal phenotypical state, respectively, when exposed for 40 days to the same time-dependent exogenous TGFβ protocol (Fig. 6). Note that the baseline model (black) does not undergo MET and remains in the mesenchymal state following the high TGFβ dose. To quantify the accuracy of the reconstruction and prediction of the virtual cell EMT and MET dynamics, we calculate the RMSD error between the virtual cell dynamics and the forecasting system predictions during the entire simulation.

The data-assimilation reconstruction for the three virtual cells (blue), and the comparison with predictions without data-assimilation (red), are shown in Fig. 7, for which the truth system (black line) represents the dynamics of the three MET-prone virtual cells described above. We find that the approach in general accurately reconstructs each cell's dynamics for the first 20 days (Fig. 7 *A–C*), during the initial EMT progression, as quantified by the low RMSD error (Fig. 7 *D–F*). However, at day 20, the reconstruction and truth systems diverge, as the baseline model fails to reproduce the MET process. This ultimately results in a failure to predict the final phenotypic state, except for the virtual cell (cell 7), which ultimately returns to the mesenchymal state. Thus, these initial simulations illustrate that the forecasting system fails to reconstruct and predict the dynamics of phenotypically different virtual cells, in particular in the context of model error associated with the MET process.

Parameter estimation in the data-assimilation reconstruction improves predictive accuracy

We next sought to improve the accuracy of our data-assimilation approach, and we hypothesized that incorporating parameter estimation into the state space of the forecasting model can account for model error and parameter uncertainty and ultimately improve predictive accuracy. As described in the section "methods," a single parameter is reconstructed during the data-assimilation algorithm. We initially chose the mRNA snail degradation rate, *kds*, as it was highlighted as a critical parameter for the MET-prone cell sub-population. With this inclusion, we find that the data-assimilation approach successfully reconstructs both the EMT and MET dynamics of all three virtual cells (Fig. 8), with a reduced RMSD error throughout the simulation. However, interestingly, despite the accurate prediction of the phenotype, the reconstruction does not predict the true value of the *kds* parameter for all cases (Fig. 8 *G–I*).

In Fig. 9, we next perform the same analysis for the entire sub-population of MET-prone cells and predict the virtual cell phenotypical distribution for comparison with the true sub-population distribution, which is shown in Fig. 5 *C*. Similar to the results shown in Fig. 7, the forecasting system completely fails to predict the phenotypical distribution of the cell population in the absence of data-assimilation corrections (Fig. 9 *A*). The data-assimilation reconstruction without parameter estimation reconstructs

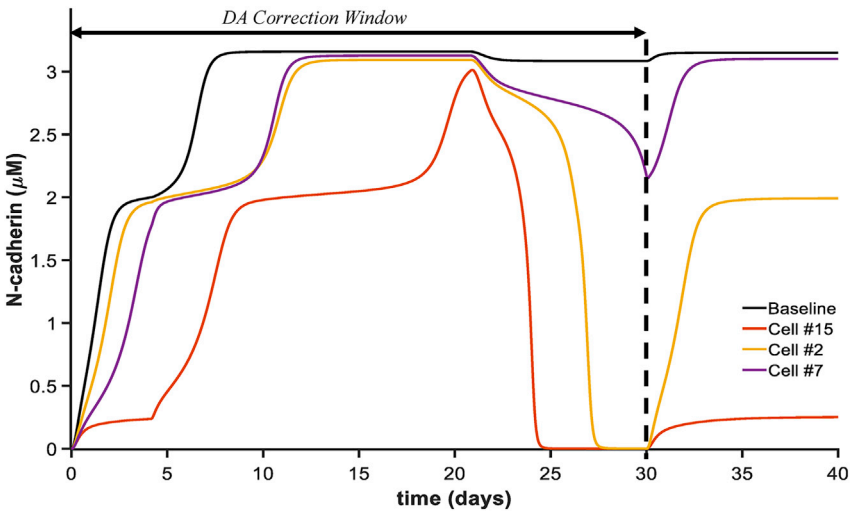

FIGURE 6 Representative EMT and MET dynamics of the MET-prone sub-population. N-cadherin expression of three MET-prone cells (15, orange; 2, yellow; and 7, purple) and the baseline model (black) are shown during the time-dependent exogenous TGFβ protocol. Note all three MET-prone cells exhibit the full EMT progression but differing MET kinetics result in different final phenotypical states at the end of the protocol. The data-assimilation (DA) correction window, in which the DA reconstruction iteratively estimates the cell state, from day 0–30 is indicated. Predictions from day 30–40 do not include DA reconstruction. To see this figure in color, go online.

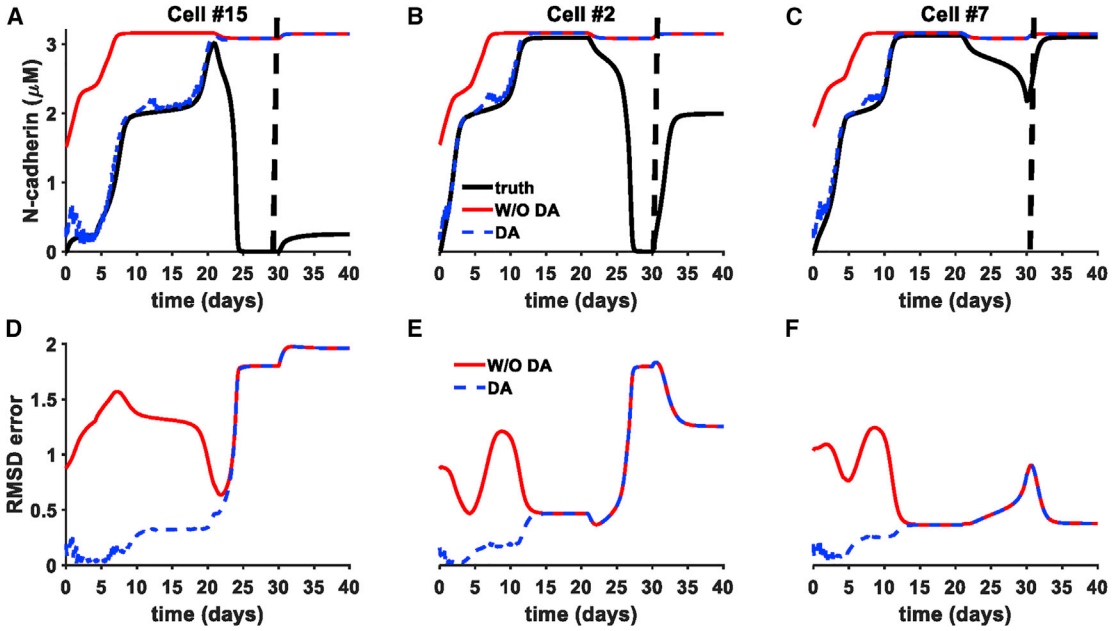

FIGURE 7 DA without parameter estimation fails to the reconstruct MET dynamics for physiological model error. (*A–C*) The true N-cadherin expression for the virtual cells (black), DA ensemble mean (dashed blue), and ensemble mean without DA corrections (red) are shown during the time-dependent TGFβ protocol. The end of the DA correction window at day 30 is denoted with a vertical dashed black line. (*D–F*) RMSD error with (blue dashed line) and without (red) DA. Parameters: truth system, parameter set from MET-prone cells 15, 2, and 7; forecasting system, baseline parameter set. To see this figure in color, go online.

fairly well the initial EMT process but not MET and subsequent final phenotype (Fig. 9 *B*). Importantly, the forecasting system successfully reconstructs the EMT and MET dynamics of the MET-prone cell sub-population when data-assimilation reconstructions incorporate parameter estimation of the critical parameter mRNA snail1 degradation (Fig. 9 *C*). Comparison of the predicted final phenotype with the truth yields accurate predictions in 95% of the MET-prone sub-population (Fig. 9 *D*), a significant improvement over the approaches without data-assimilation or parameter estimation.

We next investigate the changes in predictive accuracy for different parameters incorporated into the data-assimilation reconstruction, with the hypothesis that parameters that greatly deviate from the baseline value within the sub-population (i.e., critical parameters) will improve predictive accuracy, more so than non-critical parameters. For each parameter, we simulate the entire MET-prone sub-population response to the time-varying TGFβ dose and quantify the accuracy of the final phenotypic state prediction (Fig. 10). Note that critical parameters are denoted with an asterisk. We first find that incorporating *any* parameter estimation into the reconstruction results in greater accuracy, compared with data-assimilation without parameter estimation (dashed gray line). In general, there is variability between the predictive accuracy for different parameters, in particular the ability to accurately predict the epithelial and partial EMT states. However, several parameter choices result in correct predictions for nearly all of the 170 MET-prone cells, many of which are the critical parameters.

Critical parameter estimation in data-assimilation reconstruction

The previous results illustrate that, in general, predictive accuracy improved with reconstruction with critical parameters. Summarizing these findings for the MET-prone sub-population, the proportion of correct predictions was significantly greater for reconstructions with critical parameters, compared with non-critical parameters (Fig. 11 *A*). Further, we compare to what extent each parameter deviated from the baseline model (i.e., how critical the parameter is) with the percentage of correct predictions (Fig. 11 *B*). We find that, although there are counter-examples (i.e., high accuracy for non-critical parameters and lower accuracy for critical parameters), there is a generally positive albeit moderate correlation (Pearson correlation coefficient $r = 0.44$) between the predictive accuracy and the parametric deviation from baseline. That is, the greater the deviation of the parameter from the baseline model within the sub-population, there is a trend for greater predictive accuracy when incorporating that parameter into the data-assimilation reconstruction.

The prior analysis illustrates that incorporating parameter estimation into the reconstruction dramatically improves prediction accuracy, yet Fig. 8 *G–I* also illustrates that parameter estimation does not necessarily predict the true

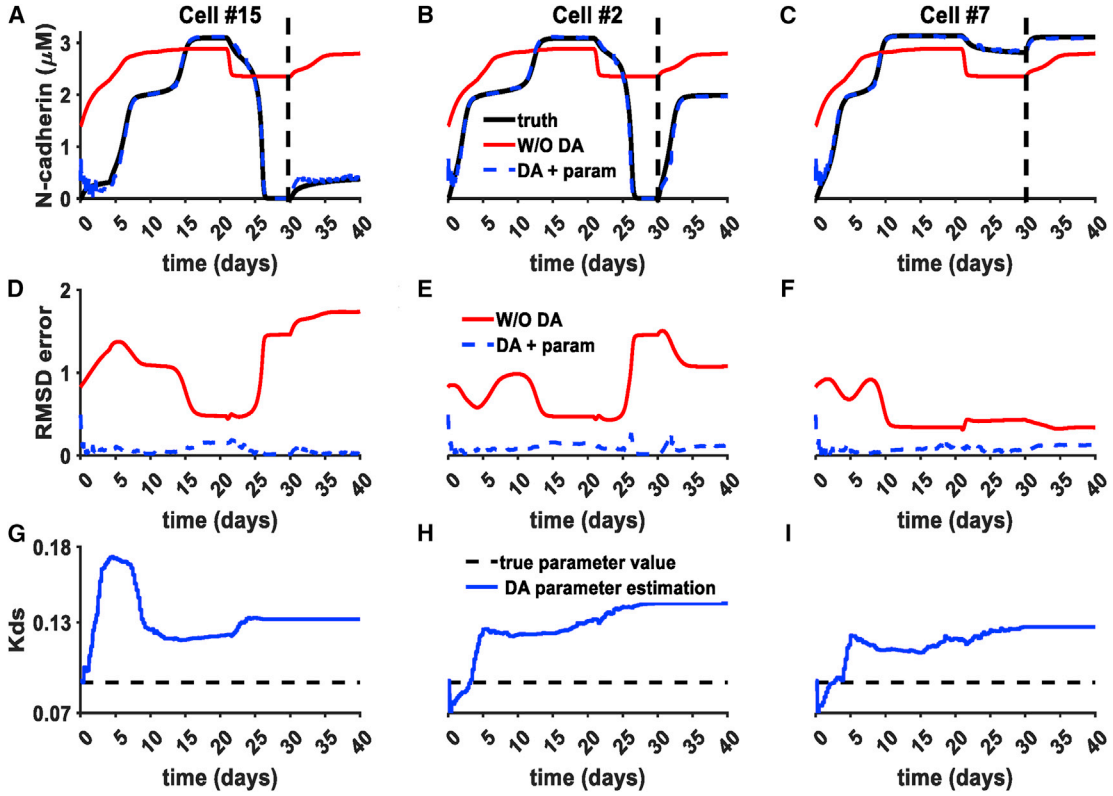

FIGURE 8 DA with parameter estimation successfully reconstructs EMT and MET dynamics for physiological model error. (*A–C*) The true N-cadherin expression for the virtual cells (black), DA ensemble mean (dashed blue), and ensemble mean without DA corrections (red) are shown during the time-dependent TGFβ protocol. The end of the DA correction window at day 30 is denoted with a vertical dashed black line. (*D–F*) RMSD error with (blue dashed line) and without (red) DA. (*G–I*) DA estimate of snail mRNA degradation parameter (*kds*, blue) and true *kds* value (dashed black) are shown. Note that the true *kds* value for the three cells are slightly different but similar. Parameters: truth system, parameter set from MET-prone cells 15, 2, and 7; forecasting system, baseline parameter set. Estimated parameter: *kds*. To see this figure in color, go online.

value of the parameter estimated. For all simulations with parameter estimation across the entire MET-prone sub-population, we analyze the error of the estimated parameter and the true parameter value to assess if there is a relationship between this error and final phenotype predictive accuracy (Fig. 12). Histograms illustrate the parameter estimation error percentage for the sub-population, showing the error after the first analysis step (6 h, blue) and at the data-assimilation correct window (30 days, orange). Interestingly, the error tends to be lower after the first analysis step compared with the correction window end, and this is true for both the parameter with the worst ($k3$, the micro-RNA 34 production rate) and the best (*kds*, the snail1 RNA degradation rate) prediction accuracy (Fig. 12 *A* and *B*). At the end of the correction window, the error is higher for the worst-performing parameter, compared with the best performing, although the error for the best performer is substantial with a mean near 50%. A scatter plot of the average of the correction window end parameter estimation error magnitude and the final phenotypical predictive accuracy is shown in Fig. 12 *C*. Across all parameters, there is a generally negative correlation ($r = -0.43$) between the predictive accuracy and average parameter estimation error.

Perturbations to promote or suppress EMT dynamics

The data-assimilation approach has successfully reconstructed and predicted the EMT and MET dynamics of a different virtual cell in the presence of time-varying TGFβ doses within the MET-prone sub-population. For a final series of experiments, we investigate the accuracy of predicting the response to a specific perturbation that either promotes or suppresses EMT. We modify the experimental design of the previous simulations to also incorporate a parametric perturbation at the end of the data-assimilation correction window (day 30), in which specified parameters are scaled accordingly to reproduce associated changes in EMT state: EMT is suppressed (i.e., E-state promoting) by increasing the miRNA-34 production rate, and EMT is promoted (i.e., M-state promoting) by increasing the zeb mRNA translation rate. A representative set of simulations of these perturbations is shown in Fig. 13. The same scaling factor is incorporated into forecasting system, although due to model error, the true value (from the MET-prone sub-population) and forecasting system value (the scaled baseline parameter) differ.

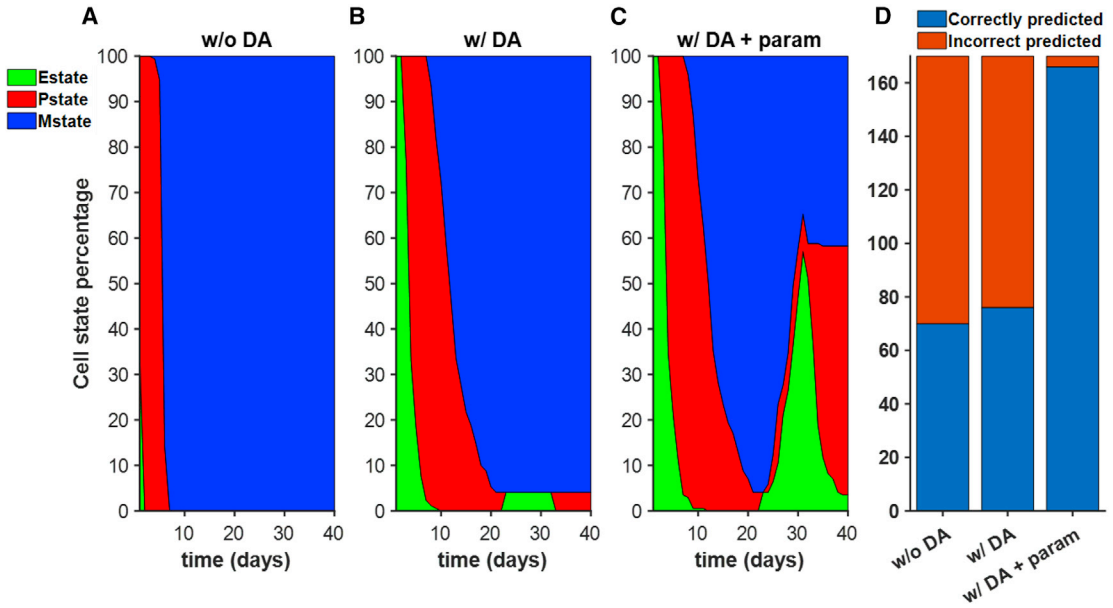

FIGURE 9 DA with parameter estimation successfully reconstructs the heterogeneous response of the MET-prone sub-population. The forecast percentage of MET-prone cells in the epithelial (E, green), partial (P, red), and mesenchymal (M, blue) state are shown as a function of time (*A*) without DA corrections, (*B*) with DA, and (*C*) with DA including parameter estimation (*kds*). The true cell state percentage is shown in Fig. 5 *C*. (*D*) Correct (blue) and incorrect (orange) predictions of the final cell states for the MET-prone sub-population (out of 170 cells). Parameters: truth system, MET-prone cell sub-population; forecasting system, baseline parameter set. Parameter estimated: $k_{d,s}$. To see this figure in color, go online.

For all cells in the sub-population, we apply a perturbation to promote either an epithelial, partial, or mesenchymal final steady state. The phenotype distribution time courses illustrate the shift in each state following the E-, P-, and M-state-promoting perturbations (Fig. 14 *A–C*). We perform the above data-assimilation reconstruction for all possible parameters, and the predictive accuracy of the final phenotype is summarized in Fig. 14 *D–F*. For all three perturbations, predictive accuracy increases with the incorporation of parameter estimation, relative to data assimilation without parameter estimation. Similar to the above analysis, estimation of critical parameters resulted in significantly greater predictive accuracy, compared with non-critical parameters, for all three perturbations. Further, the best-performing parameter was the degradation rate of snail1 mRNA (*kds*) for all three perturbations, and its predictive accuracy is nearly perfect for all perturbations as well. Additionally, predictive accuracy with critical parameters is similar for all three perturbations.

We also consider how estimations of the parameters that are perturbed affect reconstruction accuracy. The P-state-promoting perturbation parameter (*kdz*) is a critical parameter, while the E- and M-state-promoting perturbation parameters (*k*34 and *kz*, respectively) are not. Interestingly, the accuracy of the E- and P-state-promoting perturbation reconstructions was not improved, relative to the non-critical parameter and critical parameter averages, respectively, while, for the M-state-promoting perturbation parameter, the accuracy was improved relative to the non-critical parameter average. Thus, surprisingly, overall estimating the perturbation parameter did not consistently improve prediction accuracy.

Predictive accuracy of the perturbation response for multiple sub-populations

Finally, we perform the above analysis on the three other identified sub-populations (EMT prone, EMT resistant, and P-M resistant), in the absence or presence of one of the aforementioned perturbations. The predictive accuracy of the final phenotype is shown without data assimilation, with data assimilation but no parameter estimation, with data assimilation incorporating estimation of either non-critical or critical parameters, and the best performer (Fig. 15). Note that the accuracy of the MET-prone sub-population, shown above in Figs. 11 *A* and 14 *D–F*, is presented here for comparison. For the EMT-prone and P-M-resistant cell sub-populations, a P-state-promoting perturbation was not applied, since the baseline of these two sub-populations displayed a large proportion of the sub-population in the partial state at day 30 (Fig. S1).

Importantly, for all sub-populations and in the presence or absence of perturbations, predictive accuracy was increased for reconstruction with parameter estimation, compared with data-assimilation without parameter estimation. Interestingly, while predictive accuracy was similar for the different perturbations for the MET-prone sub-population, this was not the case for all sub-populations. Specifically, accuracy was generally higher predicting the phenotype following P- or M-state-promoting perturbations in the

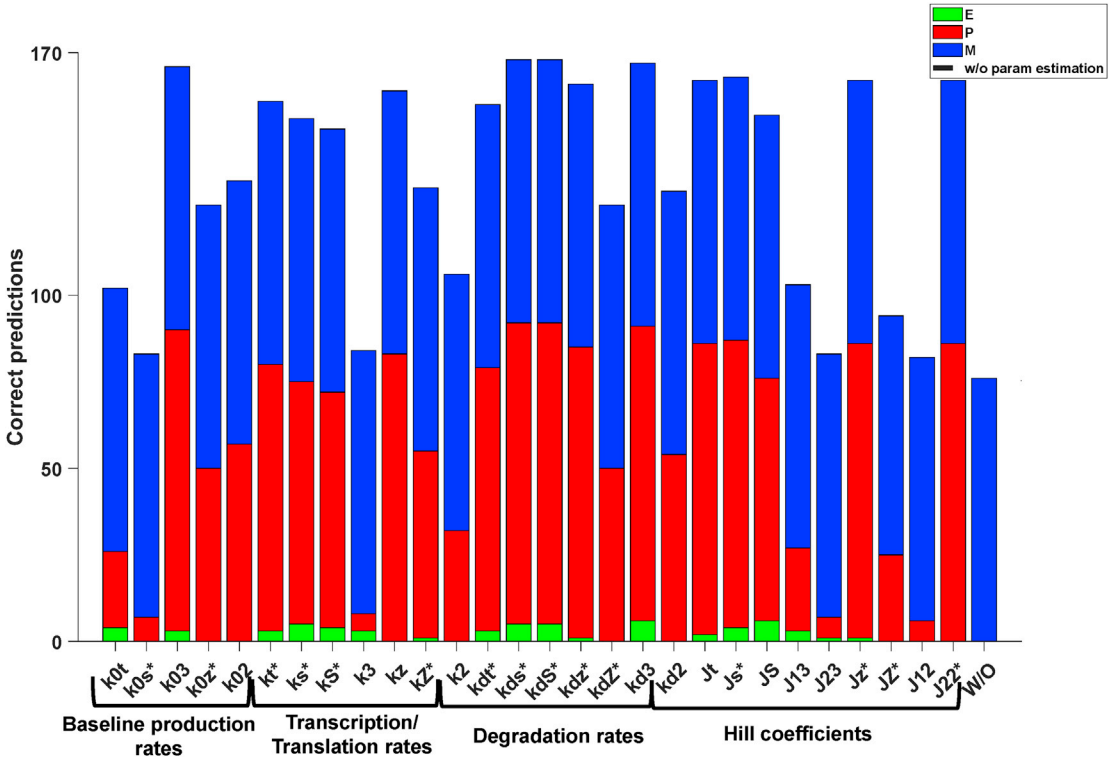

FIGURE 10 Predictive accuracy of DA approach depends on the estimated parameter. Bar plot displays the number and phenotype of accurate final phenotypical states predicted by the DA algorithm out of the 170 cell MET-prone sub-population for different parameters estimated. Parameters: truth system, MET-prone cell sub-population; forecasting system, baseline parameter set. To see this figure in color, go online.

EMT-resistant sub-population, compared with E-state promoting or no perturbation, although these differences were more pronounced for reconstructions using non-critical parameters (Fig. 15, third column). Interestingly, for the P-M-resistant sub-population, accuracy of the data-assimilation reconstruction without parameter estimation was worse than predictions without data assimilation at all (Fig. 15, fourth column). Most importantly, for all cases, predictive accuracy was high (above 90%) for the data-assimilation reconstruction with a critical parameter, with the best performer near perfect in all cases. We also importantly note that the best performer was always a critical parameter for each sub-population (which differ between the sub-populations as shown in Fig. 3). A list of the top five best performers for each of the 14 cases (different sub-populations and presence/absence of perturbations) is shown in Table S4.

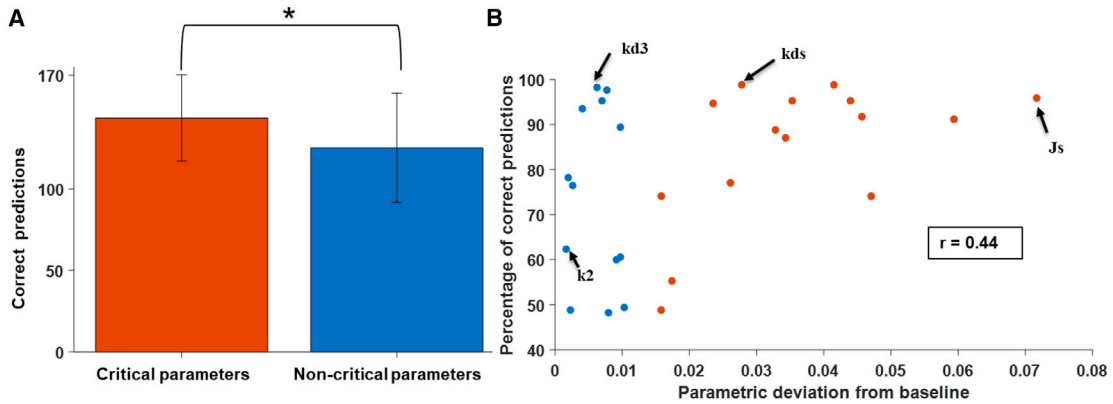

FIGURE 11 Estimation of critical parameters improves predictive accuracy. (*A*) Bar plot displays the mean number of correct final phenotype predictions for critical and non-critical parameters. Error bars represent the standard deviation. The proportion of critical parameter predictions is significantly larger than for non-critical parameters ($*p < 0.05$). (*B*) Scatter plot of the percentage of correct predictions and the parametric deviation from baseline (shown for each parameter). Prediction accuracy generally positively correlated ($r = 0.44$) with how much a parameter deviated from the baseline (i.e., how critical the parameter was). To see this figure in color, go online.

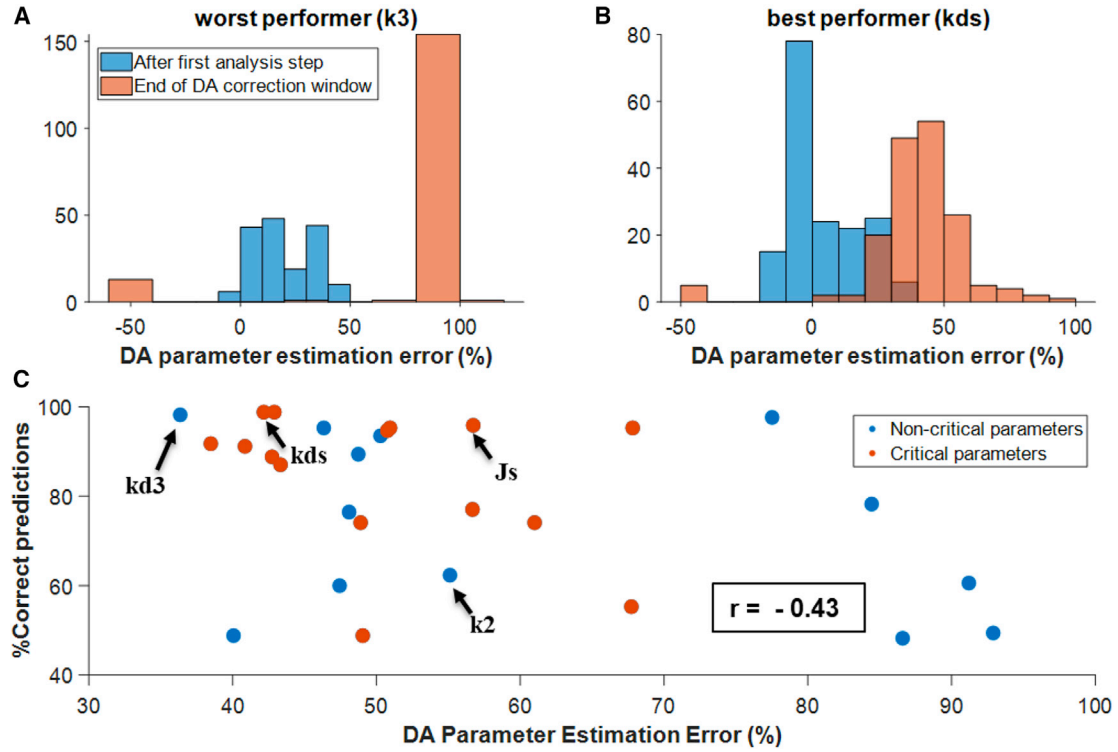

FIGURE 12 DA does not reconstruct the estimated parameter values. Histogram of the parameter estimation error at the start of the DA (blue) and at the end of the DA correction window (orange) for the (*A*) worst-performing parameter (*k3*) and (*B*) best-performing parameter (*kds*). (*C*) Scatter plot of the percentage of correct final predictions and the parameter estimation error at the end of the DA corrections window. Predictive accuracy was generally negatively correlated ($r = -0.43$) with DA parameter estimation error. To see this figure in color, go online.

DISCUSSION

In this study, we emulated the heterogeneity inherently observed in epithelial cell populations by generating a large population of virtual cells with different parameter sets. We subsequently categorized the population into four sub-populations: MET prone, EMT prone, EMT resistant, and P-M resistant. Analysis of these parameter sets revealed a series of critical parameters for which each sub-population differs significantly from the baseline. Subsequently, we performed a series of computational experiments in which we successfully applied a data-assimilation approach to accurately reconstruct EMT dynamics of heterogeneous virtual cell populations in the presence of physiological model error and parameter uncertainty. We found that the predictive accuracy of our data-assimilation approach significantly improves when coupled with single parameter estimation. Furthermore, predictive accuracy of the data-assimilation approach was further enhanced by incorporating the estimation of population-specific critical parameters, which, to our knowledge, is novel and has not been previously demonstrated.

One of the main goals of this study was to test the capabilities of the data-assimilation approach to predict the response of a heterogeneous epithelial cell population to a specific parametric perturbation. To test this goal, we performed multiple data-assimilation simulations for which parametric perturbations that promoted either epithelial, partial, or mesenchymal final cell states were applied to the four different cellular sub-populations. The data-assimilation algorithm successfully reconstructed the EMT and MET dynamics of the population, while generating accurate predictions of the population-specific responses to these perturbations. As just noted, the accuracy of the data-assimilation algorithm was consistently improved when coupled with single parameter estimation. Furthermore, the estimation of a critical parameter yielded on a statistically significant improvement of the mean accuracy rate of the data-assimilation algorithm. Interestingly, the snail1 mRNA degradation rate parameter (*kds*) was common as a top performer across multiple sub-populations and perturbations.

This series of in silico numerical experiments functions as a proof of concept for computational methods generating accurate long-term predictions of heterogeneous populations to complex inputs (e.g., time-varying stimuli and time-dependent parametric perturbations). It is noteworthy that to maximize the accuracy of computational predictions of population dynamics, it is an effective strategy to estimate a single population-specific critical parameter. Interestingly, one might speculate that estimating all parameters simultaneously would result in the greatest predictive accuracy.

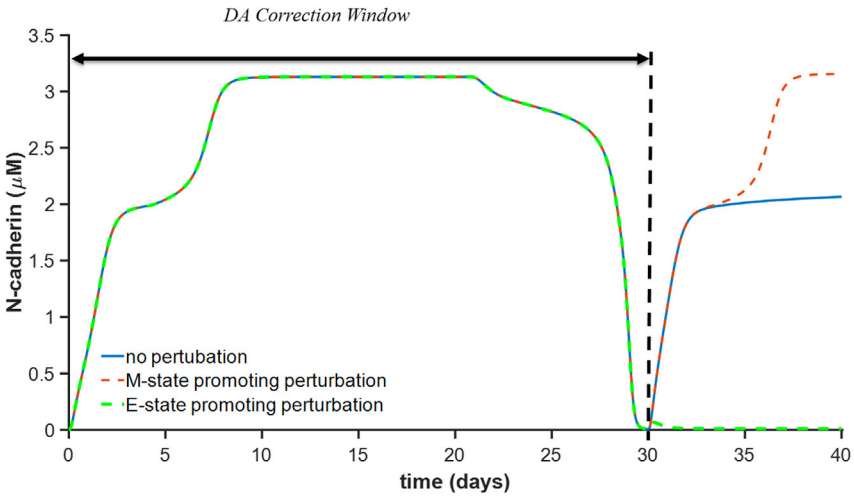

FIGURE 13 Parametric perturbations promote transitions to specific final phenotypical states. N-cadherin expression is shown for no perturbation (blue), and an M- and E-state-promoting perturbation (dashed red and green, respectively) during the time-dependent TGFβ protocol. The perturbation is applied at the end of the DA correction window (day 30, dashed black line). The E-state-promoting perturbation is increasing the miRNA-34 production rate $k34$ (scaling factor of 3). The M-state-promoting perturbation is increasing the zeb mRNA translation rate kz (scaling factor of 1.7). To see this figure in color, go online.

However, perhaps counter-intuitively, this approach not only does not increase predictive accuracy of the cell state but also does not reconstruct the true parameter values either (Fig. S2), reinforcing the hypothesis that targeted parameter estimation can offer greater improvement in computational predictive algorithms. Additionally, as data assimilation is specifically based on observations that are a subset of the full system space, applying this approach to virtual cells inherently provides insights into what cell signaling network component(s) (i.e., what is observed) yield useful information about the overall system state. These in turn motivate potentially new experiments designed to measure specific cellular components or properties. Here, we note that the specific form of the observation (i.e., the ratio of E-cadherin and ZEB) utilized in our study was motivated by a recent dual-reporter sensor that incorporates information from both stages of the two-stage EMT progression (42). However, additional simulations in which different observations are utilized, in particular ratiometric measurements involving only the first or only the second stage of the EMT progression, demonstrate less accurate and noisier reconstruction (Fig. S3).

In our previous work (19), we tested the capabilities of a data-assimilation approach to reconstruct and forecast the EMT dynamics of a single virtual epithelial cell under different degrees of parameter uncertainty and model error. The current study builds upon our previous single-cell data-assimilation simulations in several significant manners to expand the scope and applications of this approach, specifically reconstructing and predicting heterogeneous population EMT dynamics in the presence of 1) physiological levels of parameter uncertainty, 2) time-dependent stimuli, and 3) parametric perturbations that alter cell phenotype. The simultaneous expression of multiple EMT cell states by a cell population observed in numerical experiments mirrors the phenotypical heterogeneity observed in in vitro and in vivo cell population studies (15,17,43–45). Furthermore, the presence of cell sub-populations with different genetic profiles that influence their EMT dynamic trajectory resembles the variance in plasticity and stemness observed in epithelial cell populations in the setting of pathological conditions such as carcinogenic tumors (43,46). This genetic and phenotypic heterogeneity has been hypothesized to be one of the main contributors to the generation of diverse tumor cell populations, thus enhancing cancer aggressiveness and therapy resistance (47,48). Additionally, the time-dependent TGFβ and perturbation protocol applied in our computational study emulates recent in vitro investigations of EMT dynamics (15,49). We note that our simulations predict that suppression or enhancement of zeb mRNA promotes either the epithelial or mesenchymal state, respectively, consistent with the cell population response observed in in vitro studies (50,51). Similarly, the EMT-suppressing effects of miR34 overexpression have also been observed in vitro (52), consistent with model predictions.

Our study illustrates how a data-assimilation approach can be viable tool for the reconstruction and prediction of dynamical responses of heterogeneous cell populations in the presence of biochemical perturbations. This is particularly useful for the study of cellular dynamics in pathological conditions such as tumorigenesis and cancer metastasis, as phenotypic heterogeneity has been observed to be one of the main drivers of cancer aggressiveness and therapy resistance (47,48). Combined with the appropriate experimental measurements, a data-assimilation approach can offer the possibility of improving a patient-specific dynamical prediction of these pathological cellular processes and the response to EMT-promoting or -suppressing perturbations. Thus, accurate forecasting systems can be a powerful tool for the development of potential therapeutic strategies. Data-assimilation approaches have been traditionally used to reconstruct high-dimensional systems for weather forecasting (53); however, a few prior studies have applied data-assimilation approaches to different biological systems. In particular, data-assimilation

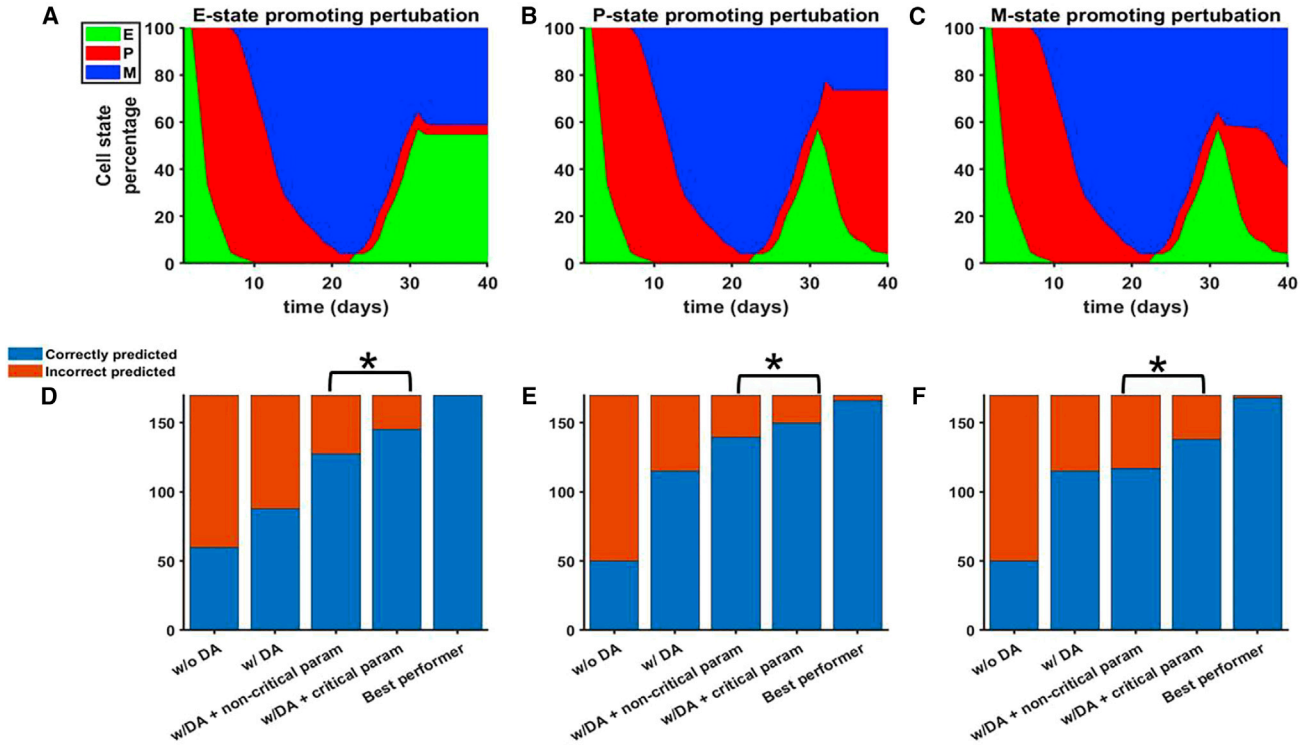

FIGURE 14 DA with parameter estimation successfully predicts the population response of MET-prone cells to specific phenotypical-state-promoting perturbations. The forecast percentages of MET-prone cells in the epithelial (E, blue), partial (P, red), and mesenchymal (M, blue) state are shown as a function of time in the presence of (*A*) an epithelial-state-promoting perturbation ($k34$ scaling factor of 3), (*B*) partial EMT-state-promoting perturbation (kdz scaling factor of 1.8), and (*C*) a mesenchymal-state-promoting perturbation (kz scaling factor of 1.7). (*D–F*) Bar plots of the number of correct (blue) and incorrect (orange) predicted final cell states for the MET-prone sub-population, out of 170 cells, reconstruction without DA, with DA (without parameter estimation), the mean for DA with non-critical parameter estimation, the mean for DA with critical parameter estimation, and the best-performing parameter. For all three perturbations, predictive accuracy is significantly greater for critical parameters, compared with non-critical parameters. ($*p < 0.05$). Parameters: truth system, MET-prone cell sub-population; forecasting system, baseline parameter set. Parameter estimated: *kds*. To see this figure in color, go online.

approaches have been mainly used in a biological context for the reconstructing of excitable cell dynamics with oscillatory and bursting behavior for various levels of scale and complexity. Ullah and Schiff applied data assimilation to reconstruct unobserved state variables in small neural networks (34,35). A data-assimilation approach was used for reconstructing the unobserved dynamical behavior of single cardiac cells, such as intracellular ionic concentrations (54). Furthermore, a similar data-assimilation approach as used in this study, but with an ensemble Kalman filter that integrates spatial components, was used to reconstruct complex electrical rhythms in one-dimensional and three-dimensional cardiac tissues (36,37).

Bayesian parameter estimation methods are an important tool for accurately reconstructing unknown biological system parameters (55,56). We note that data assimilation employs a Bayesian inference approach to reconstruct and predict system behaviors (18), with one of the most notable differences between data assimilation with parameter estimation and other Bayesian parameter estimation methods being that data-assimilation estimates (and ultimately applies corrections to) both the state space and parameter space at the same time, whereas other Bayesian parameter estimation methods generally only fit the model parameters. Additionally, in general, Bayesian parameter estimation methods are performed post hoc and require critical assumptions in the form of the prior probability distribution. The ensemble form of the data-assimilation algorithm provides a framework for generating an adequate prior probability distribution and estimating the uncertainty in model predictions. Specifically, the ensemble Kalman filter generates a prior distribution using a mechanistic model to generate an ensemble of forecasts and using the forecast distribution as the prior distribution (performed each iteration). Further, the uncertainty can be estimated using the covariance matrix of the ensemble forecasts. Thus, the data-assimilation algorithm is both based on and builds upon Bayesian parameter estimation approaches.

As we show here, coupling the reconstruction algorithm for the state space with parameter estimation is a useful tool for improving data-assimilation predictions. A general approach to implement parameter estimation for a data-assimilation algorithm usually includes augmenting the state space of the forecasting system with the selected parameters that will be set to be estimated. However, we note that previous work using data-assimilation with parameter estimation

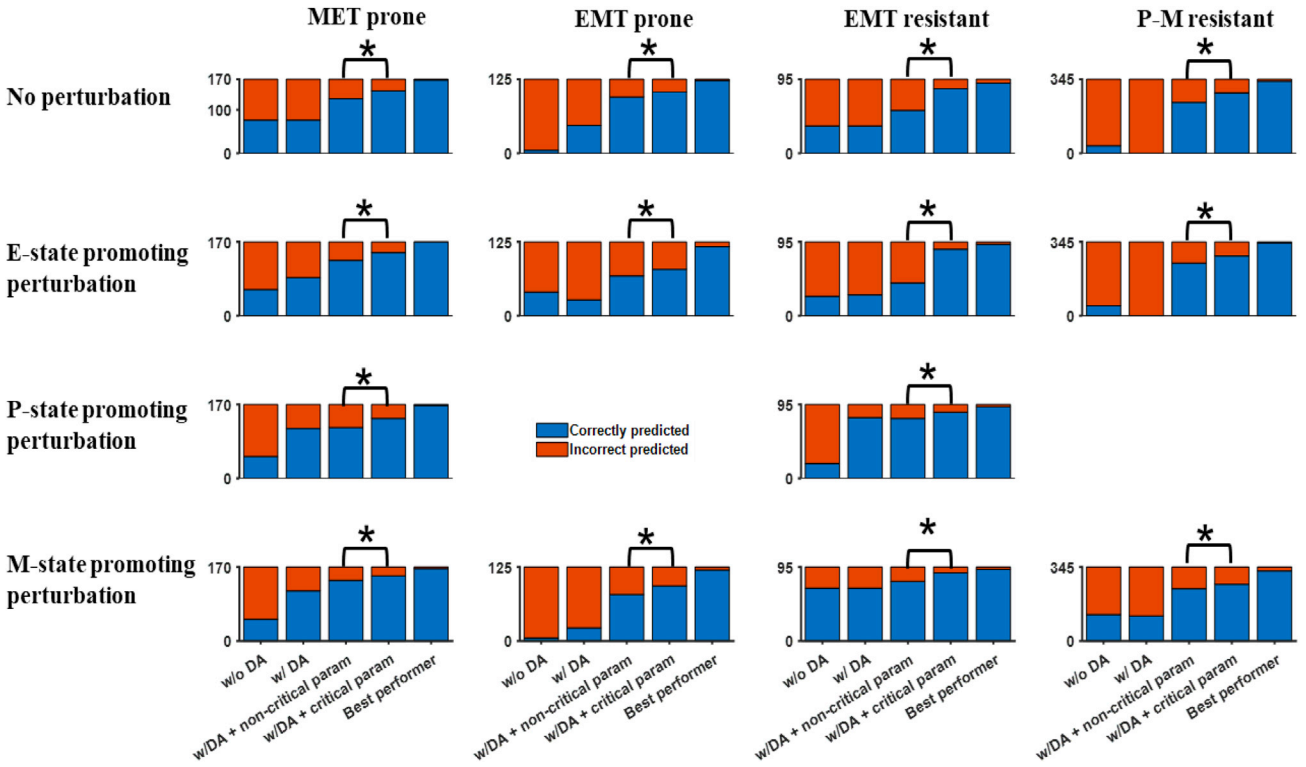

FIGURE 15 DA with parameter estimation successfully predicts the response of multiple sub-populations to specific phenotypical-state-promoting perturbations. Each panel shows bar plots of the number of correct (blue) and incorrect (orange) predicted final cell states for the MET-prone sub-population, out of 170 cells, reconstruction without DA, with DA (without parameter estimation), the mean for DA with non-critical parameter estimation, the mean for DA with critical parameter estimation, and the best-performing parameter. Columns 1–4 show results for the four different cell sub-populations: MET prone, EMT prone, EMT resistant, and P-M resistant cells. Rows 1–4 correspond with no perturbation, and perturbations promoting the E, P, and M states, respectively. For all perturbations, predictive accuracy is significantly greater for critical parameters, compared with non-critical parameters. ($*p < 0.05$). To see this figure in color, go online.

focused on the reconstruction of the unknown parameters in the presence of varying degrees of model error. Some examples of these studies include the work by Moye et al., in which they apply this expanded data assimilation to improve the estimates of both neural cell states and model parameters for different types of neuron spiking bifurcation behavior (38). A similar data-assimilation approach was used to reconstruct unknown parameters from phosphotyrosine-dependent signaling networks in the epidermal growth factor receptor (EGFR) pathways (57). Furthermore, coupling between state space corrections and parameter estimation has been applied to reconstruct the parameter and state space of blood glucose levels (39), mammalian sleep dynamics (40), and neuron signaling and neuronal networks (58–60). Importantly, we note that, in contrast to these studies, our study does not typically result in accurate prediction of the estimated parameters, but rather incorporating parameter estimation provides additional degrees of freedom to the data-assimilation reconstruction. Indeed, this result is perhaps expected, with hindsight, as successful prediction of the estimated parameter would still result in model error arising from the other 27 out of 28 unknown parameters. A similar approach known as stochastic model parameter selection was applied by one of the authors of this study in Marcotte et al. (25), in which a subset of parameters are drawn from a random distribution for each ensemble and each assimilation interval; this approach similarly found that the accuracy of the data-assimilation reconstruction of the state space was improved, without accurate reconstruction of model parameters. One hypothesis for how these results are obtained is that our approach takes advantage of the model's ability to produce constrained predictions with loosely constrained parameter sets. Sethna and coworkers refer to this phenomenon as "sloppiness" in a modeling framework, which in practice means that a region of the parameter space can reconstruct the same system dynamics. It has been suggested that this property can be advantageous for the generation of predictions of the state space of a biological system, but not for the reconstruction of the parameter space (61–63). To our knowledge, the integration of parameter estimation into the data-assimilation reconstruction to improve the state prediction, but not the unknown parameter itself, is a novel application.

This study is a proof-of-concept demonstration of applying data assimilation to predict cell population dynamics; however, there are several key limitations to be addressed in

future studies. At present, the cell population simulations focus on the core regulatory biochemical signaling pathways and do not integrate spatial and multicellular interactions occurring within a tissue during EMT. The challenge of model development of the spatial interactions during the EMT process is an area of ongoing work within our laboratory (64,65) and others (66–68). Following the methods describe by Hunt et al. (18), we plan to extend the approach demonstrated here to account for spatial localization and interacting spatial dynamics in multicellular tissues in the future. Additionally, our work utilized a single signaling network (Tian et al. model) to represent the core regulatory pathway of TGFβ-induced EMT. The mathematical relationships proposed by this model are based on key experimental findings of the interactions of critical microRNAs and transcription factors regulating the EMT process (20); however, other EMT regulation signaling pathways, such as Wnt and β-catenin signaling (69,70), are not accounted for. For future work, our approach is highly generalizable and can be applied to other models of EMT (44,45,71).

CONCLUSIONS

In this computational study, we use a data-assimilation approach to reconstruct and predict the responses of a phenotypically heterogeneous population of epithelial cells to a time-dependent EMT-inducing stimulus and EMT-suppressing or -promoting perturbations. The use of a data-assimilation approach incorporating the estimation of population-specific critical parameters greatly increased the predictive accuracy, facilitating the prediction of the responses to biochemical perturbations in real time, with future applications to cell diagnostics and patient-specific therapies.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2022.07.014.

AUTHOR CONTRIBUTIONS

M.J.M., M.J.H., E.M.C., C.A.L., and S.H.W. designed the research. M.J.M. carried out all simulations and analyzed the data. M.J.M., M.J.H., E.M.C., C.A.L., and S.H.W. wrote the article.

ACKNOWLEDGMENTS

This work was supported through funding from the National Institutes of Health/National Institute of General Medical Sciences R01GM122855 (S.H.W., C.A.L.) and the National Science Foundation DCSD-2011280 (E.M.C., M.J.H.).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Radisky, D. C. 2005. Epithelial-mesenchymal transition. *J. Cell Sci.* 118:4325–4326.
2. Thiery, J. P., H. Acloque, …, M. A. Nieto. 2009. Epithelial-mesenchymal transitions in development and disease. *Cell.* 139:871–890.
3. Brabletz, T., R. Kalluri, …, R. A. Weinberg. 2018. EMT in cancer. *Nat. Rev. Cancer.* 18:128–134.
4. Kalluri, R., and E. G. Neilson. 2003. Epithelial-mesenchymal transition and its implications for fibrosis. *J. Clin. Invest.* 112:1776–1784.
5. Miettinen, P. J., R. Ebner, …, R. Derynck. 1994. TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J. Cell Biol.* 127:2021–2036.
6. Xu, J., S. Lamouille, and R. Derynck. 2009. TGF-β-induced epithelial to mesenchymal transition. *Cell Res.* 19:156–172.
7. Griggs, L. A., N. T. Hassan, …, C. A. Lemmon. 2017. Fibronectin fibrils regulate TGF-β1-induced epithelial-mesenchymal transition. *Matrix Biol.* 60-61:157–175.
8. Scott, L. E., S. H. Weinberg, and C. A. Lemmon. 2019. Mechanochemical signaling of the extracellular matrix in epithelial-mesenchymal transition. *Front. Cell Dev. Biol.* 7:135.
9. Jolly, M. K., S. A. Mani, and H. Levine. 2018. Hybrid epithelial/mesenchymal phenotype (s): the 'fittest'for metastasis? *Biochim. Biophys. Acta. Rev. Cancer.* 1870:151–157.
10. Liao, T.-T., and M.-H. Yang. 2020. Hybrid epithelial/mesenchymal state in cancer metastasis: clinical significance and regulatory mechanisms. *Cells.* 9:623.
11. Saxena, K., M. K. Jolly, and K. Balamurugan. 2020. Hypoxia, partial EMT and collective migration: emerging culprits in metastasis. *Transl. Oncol.* 13:100845.
12. Deshiere, A., E. Duchemin-Pelletier, …, O. Filhol. 2013. Unbalanced expression of CK2 kinase subunits is sufficient to drive epithelial-to-mesenchymal transition by Snail1 induction. *Oncogene.* 32:1373–1383.
13. Taube, J. H., J. I. Herschkowitz, …, S. A. Mani. 2010. Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proc. Natl. Acad. Sci. USA.* 107:15449–15454.
14. Hesling, C., L. Fattet, …, R. Rimokh. 2011. Antagonistic regulation of EMT by TIF1γ and Smad4 in mammary epithelial cells. *EMBO Rep.* 12:665–672.
15. Karacosta, L. G., B. Anchang, …, S. K. Plevritis. 2019. Mapping lung cancer epithelial-mesenchymal transition states and trajectories with single-cell resolution. *Nat. Commun.* 10:5587–5615.
16. Zhang, J., X.-J. Tian, …, J. Xing. 2014. TGF-β–induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci. Signal.* 7:ra91.
17. Hirway, S. U., N. T. Hassan, …, S. H. Weinberg. 2021. Immunofluorescence image feature analysis and phenotype scoring pipeline for distinguishing epithelial–mesenchymal transition. *Microsc. Microanal.* 27:849–859.
18. Hunt, B. R., E. J. Kostelich, and I. Szunyogh. 2007. Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Phys. Nonlinear Phenom.* 230:112–126.
19. Mendez, M. J., M. J. Hoffman, …, S. H. Weinberg. 2020. Cell fate forecasting: a data-assimilation approach to predict epithelial-mesenchymal transition. *Biophys. J.* 118:1749–1768.
20. Tian, X.-J., H. Zhang, and J. Xing. 2013. Coupled reversible and irreversible bistable switches underlying TGFβ-induced epithelial to mesenchymal transition. *Biophys. J.* 105:1079–1089.
21. Houtekamer, P. L., and F. Zhang. 2016. Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* 144:4489–4532.
22. Kalnay, E., H. Li, …, J. Ballabrera-Poy. 2007. 4-D-Var or ensemble Kalman filter? *Tellus Dyn. Meteorol. Oceanogr.* 59:758–773. https://doi.org/10.1111/j.1600-0870.2007.00261.x.

23. Szunyogh, I., E. J. Kostelich, …, J. A. Yorke. 2008. A local ensemble transform Kalman filter data assimilation system for the NCEP global model. *Tellus Dyn. Meteorol. Oceanogr.* 60:113–130. https://doi.org/10.1111/j.1600-0870.2007.00274.x.

24. Miyoshi, T., and M. Kunii. 2011. The local ensemble transform Kalman filter with the weather research and forecasting model: experiments with real observations. *Pure Appl. Geophys.* 169:321–333. http://link.springer.com/article/10.1007/s00024-011-0373-4.

25. Marcotte, C. D., F. H. Fenton, …, E. M. Cherry. 2021. Robust data assimilation with noise: applications to cardiac dynamics. *Chaos.* 31:013118.

26. Annan, J. D., and J. C. Hargreaves. 2004. Efficient parameter estimation for a highly chaotic system. *Tellus.* 56:520–526.

27. Dee, D. P., and A. M. Da Silva. 1998. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.* 124:269–295.

28. Zupanski, D., and M. Zupanski. 2006. Model error estimation employing an ensemble data assimilation approach. *Mon. Weather Rev.* 134:1337–1354.

29. Koyama, H., and M. Watanabe. 2010. Reducing forecast errors due to model imperfections using ensemble Kalman filtering. *Mon. Weather Rev.* 138:3316–3332.

30. Kostelich, E. J., Y. Kuang, …, M. C. Preul. 2011. Accurate state estimation from uncertain data and models: an application of data assimilation to mathematical models of human brain tumors. *Biol. Direct.* 6:64.

31. Hamilton, F., T. Berry, …, T. Sauer. 2013. Real-time tracking of neuronal network structure using data assimilation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 88:052715.

32. Hamilton, F., J. Cressman, …, T. Sauer. 2014. Reconstructing neural dynamics using data assimilation with multiple models. *Europhys. Lett.* 107:68005.

33. Hamilton, F., T. Berry, and T. Sauer. 2018. Tracking intracellular dynamics through extracellular measurements. *PLoS One.* 13:e0205031.

34. Ullah, G., and S. J. Schiff. 2010. Assimilating seizure dynamics. *PLoS Comput. Biol.* 6:e1000776.

35. Ullah, G., and S. J. Schiff. 2009. Tracking and control of neuronal Hodgkin-Huxley dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 79:040901.

36. Hoffman, M. J., N. S. LaVigne, …, E. M. Cherry. 2016. Reconstructing three-dimensional reentrant cardiac electrical wave dynamics using data assimilation. *Chaos.* 26:013107.

37. LaVigne, N. S., N. Holt, …, E. M. Cherry. 2017. Effects of model error on cardiac electrical wave state reconstruction using data assimilation. *Chaos.* 27:093911.

38. Moye, M. J., and C. O. Diekman. 2018. Data assimilation methods for neuronal state and parameter estimation. *J. Math. Neurosci.* 8:11.

39. Sedigh-Sarvestani, M., D. J. Albers, and B. J. Gluckman. 2012. Data assimilation of glucose dynamics for use in the intensive care unit. *In* 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 5437–5440.

40. Sedigh-Sarvestani, M., S. J. Schiff, and B. J. Gluckman. 2012. Reconstructing mammalian sleep dynamics with data assimilation. *PLoS Comput. Biol.* 8:e1002788.

41. Sobie, E. A. 2009. Parameter sensitivity analysis in electrophysiological models using multivariable regression. *Biophys. J.* 96:1264–1274.

42. Toneff, M. J., A. Sreekumar, …, J. M. Rosen. 2016. The Z-cad dual fluorescent sensor detects dynamic changes between the epithelial and mesenchymal cellular states. *BMC Biol.* 14:47.

43. Genna, A., A. M. Vanwynsberghe, …, C. Gilles. 2020. EMT-associated heterogeneity in circulating tumor cells: sticky friends on the road to metastasis. *Cancers.* 12:1632.

44. Jolly, M. K., C. Ward, …, S. S. Sohal. 2018. Epithelial–mesenchymal transition, a spectrum of states: role in lung development, homeostasis, and disease. *Dev. Dyn.* 247:346–358.

45. Jolly, M. K., and T. Celià-Terrassa. 2019. Dynamics of phenotypic heterogeneity associated with EMT and stemness during cancer progression. *J. Clin. Med.* 8:1542.

46. Sousa, B., A. S. Ribeiro, and J. Paredes. 2019. Heterogeneity and plasticity of breast cancer stem cells. *Adv. Exp. Med. Biol.* 1139:83–103.

47. Song, K.-A., M. J. Niederst, …, A. C. Faber. 2018. Epithelial-to-mesenchymal transition antagonizes response to targeted therapies in lung cancer by suppressing BIM. *Clin. Cancer Res.* 24:197–208.

48. Nakamichi, S., M. Seike, …, A. Gemma. 2018. Overcoming drug-tolerant cancer cell subpopulations showing AXL activation and epithelial–mesenchymal transition is critical in conquering ALK-positive lung cancer. *Oncotarget.* 9:27242–27255.

49. Krishnaswamy, S., N. Zivanovic, …, B. Bodenmiller. 2018. Learning time-varying information flow from single-cell epithelial to mesenchymal transition data. *PLoS One.* 13:e0203389.

50. Hong, T., K. Watanabe, …, X. Dai. 2015. An Ovol2-Zeb1 mutual inhibitory circuit governs bidirectional and multi-step transition between epithelial and mesenchymal states. *PLoS Comput. Biol.* 11:e1004569.

51. Wellner, U., J. Schubert, …, T. Brabletz. 2009. The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nat. Cell Biol.* 11:1487–1495.

52. Kim, N. H., H. S. Kim, …, S. J. Weiss. 2011. p53 and microRNA-34 are suppressors of canonical Wnt signaling. *Sci. Signal.* 4:ra71.

53. Ghil, M., and P. Malanotte-Rizzoli. 1991. Data assimilation in meteorology and oceanography. *Adv. Geophys.* 33:141–266.

54. Munoz, L. M., and N. F. Otani. 2013. Kalman filter based estimation of ionic concentrations and gating variables in a cardiac myocyte model. *In* Computing in Cardiology 2013. IEEE, pp. 53–56.

55. Beattie, K. A., A. P. Hill, …, G. R. Mirams. 2018. Sinusoidal voltage protocols for rapid characterisation of ion channel kinetics. *J. Physiol.* 596:1813–1828.

56. Sundnes, J., and R. Rodríguez-Cantano. 2022. A Bayesian approach to parameter estimation in cardiac mechanics. *In* Solid (Bio) Mechanics: Challenges of the Next Decade. Springer, pp. 245–256.

57. Tasaki, S., M. Nagasaki, …, S. Miyano. 2006. Modeling and estimation of dynamic EGFR pathway by data assimilation approach using time series proteomic data. *Genome Inform.* 17:226–238.

58. Meliza, C. D., M. Kostuk, …, H. D. I. Abarbanel. 2014. Estimating parameters and predicting membrane voltages with conductance-based neuron models. *Biol. Cybern.* 108:495–516.

59. Kadakia, N., E. Armstrong, …, H. D. Abarbanel. 2016. Nonlinear statistical data assimilation for HVC_{RA} neurons in the avian song system. *Biol. Cybern.* 110:417–434.

60. Wang, J., D. Breen, …, G. Cauwenberghs. 2016. Data assimilation of membrane dynamics and channel kinetics with a neuromorphic integrated circuit. *In* 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE, pp. 584–587.

61. Brown, K. S., and J. P. Sethna. 2003. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 68:021904.

62. Gutenkunst, R. N., J. J. Waterfall, …, J. P. Sethna. 2007. Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* 3:1871–1878.

63. Daniels, B. C., Y.-J. Chen, …, C. R. Myers. 2008. Sloppiness, robustness, and evolvability in systems biology. *Curr. Opin. Biotechnol.* 19:389–395.

64. Hirway, S. U., C. A. Lemmon, and S. H. Weinberg. 2021. Multicellular mechanochemical hybrid cellular Potts model of tissue formation during epithelial-mesenchymal transition. *Comput. Syst. Oncol.* 1:e1031.

65. Scott, L. E., L. A. Griggs, …, S. H. Weinberg. 2019. A predictive model of intercellular tension and cell-matrix mechanical interactions in a multicellular geometry. Preprint at bioRxiv . 701037. https://doi.org/10.1101/701037.

66. Bocci, F., L. Gearhart-Serna, …, M. K. Jolly. 2019. Toward understanding cancer stem cell heterogeneity in the tumor microenvironment. *Proc. Natl. Acad. Sci. USA.* 116:148–157.

67. Salgia, R., I. Mambetsariev, …, M. Sattler. 2018. Modeling small cell lung cancer (SCLC) biology through deterministic and stochastic mathematical models. *Oncotarget.* 9:26226–26242.

68. Metzcar, J., Y. Wang, …, P. Macklin. 2019. A review of cell-based computational modeling in cancer biology. *JCO Clin. Cancer Inform.* 3:1–13.

69. Basu, S., S. Cheriyamundath, and A. Ben-Ze'ev. 2018. Cell–cell adhesion: linking Wnt/β-catenin signaling with partial EMT and stemness traits in tumorigenesis. *F1000Research.* 7:1488.

70. Hua, K., Y. Li, …, H. Jin. 2018. Haemophilus parasuis infection disrupts adherens junctions and initializes EMT dependent on canonical Wnt/β-catenin signaling pathway. *Front. Cell. Infect. Microbiol.* 8:324.

71. Lu, M., M. K. Jolly, …, E. Ben-Jacob. 2013. MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. *Proc. Natl. Acad. Sci. USA.* 110:18144–18149.