THE UNIVERSITY OF READING

Data Assimilation Using Observers

Anne K. Griffith and Nancy K. Nichols

Numerical Analysis Report 11/94

DEPARTMENT OF MATHEMATICS

Data Assimilation Using Observers

Anne K. Griffith and Nancy K. Nichols

Numerical Analysis Report 11/94

Department of Mathematics
P.O.Box 220
University of Reading
Whiteknights
Reading
RG6 2AX
United Kingdom

Abstract

This report gives a brief introduction to data assimilation and discusses how this can be treated as an observer design problem. The particular observer design investigated here endevours to make the resulting observer system as robust as possible to perturbations in the model equations.

This observer is tested in the context of data assimilation for a simple discrete model. Issues investigated include the choice of eigenvalues to be assigned to the observer, a choice of a suitable observation matrix, and modifications for the case where observations occur less frequently.

Finally, the choice of the weighting matrix in the Cressman data assimilation scheme is compared to the feedback matrix of the observer system. This facilitates a theoretical evaluation of the Cressman scheme.

Acknowledgements

The work described in this report was carried out under an EPSRC CASE studentship with the UK Meteorological Office. We would like to thank Andrew Lorenc of the Met Office for his help and support with this work.

Contents

1 Introduction

1.1 Introduction to data assimilation

In numerical weather prediction, data assimilation is used to integrate observed data into a forecast model. The crudest method of assimilation is simply to substitute the observed values in place of the predicted values they represent. However, if the values at observation points are changed in this way, the grid-point values are no longer in agreement with each other. Data assimilation schemes therefore aim to modify the predictions so that the corrected values are consistent with both the observations and the known dynamics of the system.

Data assimilation has been widely used in various forms in meteorological and oceanographic modelling since the 1950's. The various forms use ideas from different branches of mathematics; notably probability theory, optimization and control theory. Although the problem may be formulated using different disciplines of mathematics, the resulting schemes have many common features and properties. (See [4] for an overview of different data assimilation techniques, and an extensive list of references.) In this report we examine the use of observers for data assimilation. In a complementary report [5] we examine the use of optimal control theory in data assimilation.

1.2 Data assimilation and observers

In control theory, an *observer* uses observations to drive a model state closer to the "true" state as characterised by the observations. Hence, it seems quite natural to express the data assimilation problem in terms of observers. This has been widely researched in the context of Kalman filters. The Kalman filter is an observer that seeks to drive the model state not to the observations exactly, but to the "most likely state" according to the error covariances of the observations and of the model. However, the need to recalculate these error covariances for every model timestep renders the Kalman filter a very expensive method of data assimilation. Also, the Kalman filter is not always sufficiently robust, and there are difficulties in extending the method for use with nonlinear models. Apart from the Kalman filter, data assimilation techniques have rarely evoked the theory of observers. Even so, many of the data assimilation schemes could be expressed in

terms of observers.

In Section 2, theory for developing a robust dynamic observer for a simple model is put forward, and an algorithm for implementing the observer is given. Section 3 describes experiments carried out with such an observer, using the heat equation with the "theta method" discretisation as a model. Experiment 1 investigates the best eigenvalue assignment in the development of the observer, aiming for quick convergence of the observer to the true solution. Experiment 2 looks at a suitable choice of the observation matrix C and Experiment 3 investigates the performance of the observer when the observations are infrequent. In Section 4 the Cressman scheme, a simple "successive correction" method for data assimilation, is compared to the observer method, both theoretically and practically. Section 5 summarises the conclusions drawn from the study, and gives suggestions for future work.

2 Eigenvalue assignment

We consider the following discrete linear time invariant system $\mathcal{S}$ with state $\mathbf{w}^k$ and input $\mathbf{u}^k$ at time-level t_k

$$\mathcal{S}: \qquad E\mathbf{w}^{k+1} = A\mathbf{w}^k + B\mathbf{u}^k \tag{2.1}$$

with p observations $\mathbf{y}^k$ of $\mathbf{w}^k$ given by

$$\mathbf{y}^k = C\mathbf{w}^k \tag{2.2}$$

where $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$; and B and C are full column and row rank, respectively. We assume here that the matrix E is *non-singular*. The results can, however, also be extended to generalized systems where E is singular.

Definition 1 *The system $\mathcal{S}$ (2.1) with observations (2.2) is* completely observable *if and only if for any time t_0 and any state $\mathbf{w}^0$ at time t_0 there exists a finite time t_k such that given $\mathbf{u}^j$ and $\mathbf{y}^j$ for $j = 0, 1, .., k$, the initial state $\mathbf{w}^0$ can be determined.*

See [1] Sections 4.2 and 4.7.
Necessary and sufficient conditions for the system to be completely observable are given by the following theorem, known as the Hautus condition [3]:

Theorem 1 *The system $\mathcal{S}$ (with E non-singular) is completely observable if and only if* $\forall \mu \in \mathbb{C}$
$(A - \mu E)\mathbf{v} = 0$ *and* $C\mathbf{v} = 0 \quad \Leftrightarrow \mathbf{v} = 0.$

This condition implies that if $\mathbf{v}$ is an eigenvector of $E^{-1}A$, then $C\mathbf{v} \neq 0$.

2.1 Design of a dynamic observer for $\mathcal{S}$

We suppose that system $\mathcal{S}$ is a good representation of reality, but that we do not know the initial condition $\mathbf{w}^0$. Starting from a guess $\hat{\mathbf{w}}^0$ of the initial condition, a dynamic observer uses the observations to drive the state $\hat{\mathbf{w}}^k$ to the true state $\mathbf{w}^k$ as k increases.

We form the observer as a new system $\mathcal{S}'$:

$$\mathcal{S}': \qquad E\hat{\mathbf{w}}^{k+1} = A\hat{\mathbf{w}}^k + B\mathbf{u}^k + G(\mathbf{y}^k - C\hat{\mathbf{w}}^k) \tag{2.3}$$

We want to construct the feedback matrix G so that $\hat{\mathbf{w}}^k \to \mathbf{w}^k$ as $k \to \infty$, regardless of the true initial condition $\mathbf{w}^0$, which is unknown.
Subtracting (2.3) from (2.1) and using (2.2) we have:

$$E(\mathbf{w}^{k+1} - \hat{\mathbf{w}}^{k+1}) = A(\mathbf{w}^k - \hat{\mathbf{w}}^k) - GC(\mathbf{w}^k - \hat{\mathbf{w}}^k), \tag{2.4}$$

so defining $\mathbf{e}^k = \mathbf{w}^k - \hat{\mathbf{w}}^k$, we have the error equation

$$E\mathbf{e}^{k+1} = A\mathbf{e}^k - GC\mathbf{e}^k. \tag{2.5}$$

Hence, for $\mathbf{e}^k \to 0$ as $k \to \infty$, we require that the eigenvalues of $E^{-1}(A - GC)$ have modulus less than unity.

If E is invertible (as assumed) and $\mathcal{S}$ is observable, then we can construct G to do this; in fact we can choose G to assign any eigenvalues we wish to the system $\mathcal{S}'$ [3]. Since this "inverse eigenvalue problem" is not uniquely determined [6], we have a certain amount of freedom to choose the eigenvectors as well. We can use this freedom to make the system as robust to perturbations as possible.

In [6] it is shown that for a robust system we require $\mathrm{cond}(X)$ to be as small as possible, where X is the modal matrix whose columns are the right eigenvectors corresponding to our chosen eigenvalues. Our objective, then, is for a suitable eigen*structure* assignment.

2.2 Eigenstructure assignment - theory

Eigenvalue assignment

We suppose that the set of eigenvalues we wish to assign is

$$\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}; \tag{2.6}$$

where

$$\lambda_i \in \mathbb{C}, \quad |\lambda_i| < 1, \quad \text{and} \quad \lambda_i \in \Lambda \Rightarrow \bar{\lambda} \in \Lambda \qquad \text{for} \quad i = 1, ..., n. \tag{2.7}$$

We let $D = diag[\lambda_i]$ and let X be the modal matrix of right eigenvectors of $E^{-1}(A - GC)$ and Y be the modal matrix of $E^{-T}(A^T - C^T G^T)$. Then our problem is to choose G and X to satisfy

$$(A - GC)X = EXD, \tag{2.8}$$

or, equivalently, to choose Y and G^T to satisfy

$$(A^T - C^T G^T)Y = E^T Y D. \tag{2.9}$$

For our purposes, we work with equation (2.9).

If we calculate the QR decompostion of C^T, we find that

$$C^T = [\tilde{Q}_c, Q_c] \begin{bmatrix} R_o \\ 0 \end{bmatrix}, \tag{2.10}$$

where $\tilde{Q}_c$ is $n \times p$, Q_c is $n \times (n-p)$; $[\tilde{Q}_c, Q_c]$ is orthogonal and R_o is $p \times p$ upper triangular, non singular. Substituting this into (2.9) and rearranging gives

$$\begin{pmatrix} \tilde{Q}_c^T A^T Y - \tilde{Q}_c^T E^T Y D \\ Q_c^T A^T Y - Q_c^T E^T Y D \end{pmatrix} = \begin{pmatrix} R_o G^T Y \\ 0 \end{pmatrix}, \tag{2.11}$$

from which we have

$$G^T = R_o^{-1} \tilde{Q}_c^T (A^T Y - E^T Y D) Y^{-1}, \tag{2.12}$$

$$0 = Q_c^T (A^T Y - E^T Y D). \tag{2.13}$$

Equation (2.12) is the equation for G^T for a given Y and equation (2.13) gives us a condition for choosing Y.

From (2.13) we have that for $i = 1, ..., n$

$$Q_c^T (A^T - \lambda_i E^T) \mathbf{y}_i = 0, \tag{2.14}$$

where $\mathbf{y}_i$ is the i^{th} column of Y and is the left eigenvector corresponding to eigenvalue λ_i. Therefore,

$$\mathbf{y}_i \in \mathcal{N}_i = \mathcal{N}(Q_c^T (A^T - \lambda_i E^T), \tag{2.15}$$

where $\mathcal{N}$ represents the right null space. This gives some restriction on the choice of each column $\mathbf{y}_i$ of Y, but since $\mathcal{N}_i$ has dimension p (by observability), there is still some freedom in choosing the $\mathbf{y}_i$ if $p > 1$. We can use this freedom to ensure that our selected eigenvalues are as insensitive as possible to perturbations in A, E, C and G and thus that the system is *robust*.

Eigenvalue assignment for robustness

The sensitivity of eigenvalue λ_i to perturbations in the components of A, E, C and G is given by

$$c_i = \frac{\|\mathbf{x}_i\|_2 \|E^T \mathbf{y}_i\|_2}{|\mathbf{y}_i^T E \mathbf{x}_i|}, \tag{2.16}$$

where $\mathbf{x}_i$ are the columns of X, and $\mathbf{y}_i^T$ the rows of Y^T (see [6]). If we scale $\mathbf{x}_i$ and $\mathbf{y}_i$ such that

$$\|E^T \mathbf{y}_i\|_2 = 1 \tag{2.17}$$

and

$$\left|\mathbf{y}_i^T E \mathbf{x}_i\right| = 1, \tag{2.18}$$

then to minimize c_i we must minimize $\|\mathbf{x}_i\|_2$. For the optimal conditioning we must minimize all the c_i together, and hence we must choose the columns of X to minimize

$$\nu = \sum_i c_i^2 = \sum_i \|\mathbf{x}_i\|_2^2 \equiv \|X\|_F^2, \tag{2.19}$$

where $\|.\|_F$ is the Frobenius norm.

From the scaling equations (2.17) and (2.18) we have

$$Y^T E X = I, \tag{2.20}$$

so

$$X = (Y^T E)^{-1}, \tag{2.21}$$

and hence optimizing the conditioning could equivalently be done by choosing the rows $\mathbf{y}_i^T$ of Y^T so that $\|(Y^T E)^{-1}\|_F$ is minimised. We note that

$$c_i = \frac{1}{|\cos \alpha|} \tag{2.22}$$

where α is the angle between $\mathbf{y}_i$ and $E\mathbf{x}_i$, by the scalar product rule. Therefore c_i is smallest where α is smallest, which will be where $\mathbf{y}_i$ is parallel to $E\mathbf{x}_i$. To optimize the conditioning, then, we could choose each $\mathbf{y}_i$ to be "as parallel as possible" to $E\mathbf{x}_i$. From (2.20) we have

$$\mathbf{y}_j^T E \mathbf{x}_i = 0 \quad \forall j \neq i, \tag{2.23}$$

that is, $E\mathbf{x}_i$ is orthogonal to $\mathbf{y}_j$ for all $j \neq i$. If $\mathbf{y}_i$ is to be parallel to $E\mathbf{x}_i$, it follows that $\mathbf{y}_i$ should also be orthogonal to $\mathbf{y}_j$ for $i \neq j$. A necessary condition for

optimal conditioning is therefore that the vectors $\mathbf{y}_j$ to be as close to orthogonal to each other as possible.

To summarise, our aim is to choose a set of vectors $\mathbf{y}_i$, the columns of Y so that for all $i = 1, 2, ..n$

a) $\mathbf{y}_i \in \mathcal{N}_i = \mathcal{N}(Q_c^T(A^T - \lambda_i E^T)) \quad \forall i = 1, .., n$

b) the $\mathbf{y}_i$ are linearly independent

c) the $\mathbf{y}_i$ are as close to orthogonal to each other as possible.

Condition **b)** is included, because the inverse of Y is needed for evaluating G. The set of vectors $\mathbf{y}_i$ must then be scaled so that (2.17) and (2.18) hold.

2.3 A method for eigenstructure assignment

The method described here involves choosing a set of vectors $\mathbf{y}_i$ which sastisfy conditions **a)**, **b)** and **c)** of Section 2.2, and follows a method given in [9].

Calculating the QR decompostion of $(A - \lambda_i E)Q_c$ gives

$$(A - \lambda_i E)Q_c = [\tilde{S}_i, S_i] \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \tag{2.24}$$

where $[\tilde{S}_i, S_i]$ is orthogonal and R_i is $(n-p) \times (n-p)$ upper triangular and nonsingular, $\tilde{S}_i$ is $n \times (n-p)$, $\quad S_i$ is $n \times p$. So

$$Q_c^T(A^T - \lambda_i E^T) = (R_i^T, 0) \begin{pmatrix} \tilde{S}_i^T \\ S_i^T \end{pmatrix} = R_i^T \tilde{S}_i^T \tag{2.25}$$

and hence by orthogonality of $[\tilde{S}_i, S_i]$,

$$Q_c^T(A^T - \lambda_i E^T)S_i = 0. \tag{2.26}$$

Therefore, if $\mathbf{y}_i$ is in the space spanned by the columns of S_i, then

$$Q_c^T(A^T - \lambda_i E^T)\mathbf{y}_i = 0 \tag{2.27}$$

and condition **a)** is satisfied.

We now choose any set of linearly independent left eigenvectors $\mathbf{y}_i$ satisfying condition **a)**, and modify these in turn to satisfy condition **c)**. Let

$$Y_{-i} = \{\mathbf{y}_1, ..., \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, ..., \mathbf{y}_n\}. \tag{2.28}$$

We want $\mathbf{y}_i$ to be as close to orthogonal as possible to this set. Calculating the QR decompostion gives

$$Y_{-i} = [\tilde{Z}_i, \mathbf{z}_i] \begin{bmatrix} \tilde{Y}_i \\ 0 \end{bmatrix}, \tag{2.29}$$

where $[\tilde{Z}_i, \mathbf{z}_i]$ is orthogonal, $\tilde{Y}_i$ is upper triangular and nonsingular, and $\mathbf{z}_i$ is an $n \times 1$ vector. This gives us the vector $\mathbf{z}_i$ which is orthogonal to Y_{-i}, but $\mathbf{z}_i$ may not be in S_i, which would violate condition **a**). Choosing $\mathbf{y}_i$ to be the orthogonal projection of $\mathbf{z}_i$ into S_i ensures that $\mathbf{y}_i$ is as orthogonal as possible to the set Y_{-i} whilst satisfying condition **a**). So, after normalization to ensure (2.17) holds, we take

$$\mathbf{y}_i = S_i S_i^T \mathbf{z}_i / \|E^T S_i S_i^T \mathbf{z}_i\|_2. \tag{2.30}$$

When all the columns have been modified in this way, the same procedure can then be repeated to modify the $\mathbf{y}_i$ again, until $\|(Y^T E)^{-1}\|_F$ reaches a local minimum. The feedback matrix G can then be calculated from (2.12), using the Y derived.

This method for improving the robustness of the system can not be guaranteed to converge to the minimum possible value of $\|(Y^T E)^{-1}\|_F$, but in practice it has been found to reduce its initial value significantly.

2.4 An algorithm for a robust observer

1) Calculate the QR decomposition of C^T into

$$C^T = \left[\tilde{Q}_c, Q_c\right] \begin{bmatrix} R_o \\ 0 \end{bmatrix}. \tag{2.31}$$

2) For each $i = 1, .., n$,
calculate the QR decomposition of $(A - \lambda_i E)Q_c$ into

$$(A - \lambda_i E)Q_c = \left[\tilde{S}_i, S_i\right] \begin{bmatrix} R_i \\ 0 \end{bmatrix}. \tag{2.32}$$

3) Choose columns from each of the S_i as columns of the first guess Y, in such a way that Y is invertible.

4) For $i = 1, ..., n$, modify the columns $\mathbf{y}_i$ of Y as follows:

4a) calculate the QR decomposition of $Y_{-i} = \{\mathbf{y}_1, .., \mathbf{y}_{i-1}, \mathbf{y}_{i+1}, .., \mathbf{y}_n\}$ into

$$Y_{-i} = [\tilde{Z}_i, \mathbf{z}_i] \begin{bmatrix} \tilde{Y}_i \\ 0 \end{bmatrix}. \tag{2.33}$$

4b) project the vector $\mathbf{z}_i$ into space S_i to satisfy condition **a**) and then normalize:

$$\mathbf{y}_i = S_i S_i^T \mathbf{z}_i / \| E^T S_i S_i^T \mathbf{z}_i \|_2. \tag{2.34}$$

5) repeat step 4 until $\|(Y^T E)^{-1}\|_F$ reaches a local minimum.

6) using the Y found, let the feedback matrix be G where

$$G^T = R_o^{-1} \tilde{Q}_c^T (A^T Y - E^T Y D) Y^{-1}. \tag{2.35}$$

.

3 Implementation of the method

In this section, the theory of Section 2 is tried out for a simple model, which is introduced in Section 3.1. Experiment 1 described in Section 3.2 investigates how the choice of the set of eigenvalues Λ affects the results. In Experiment 2 (Section 3.3), different forms for the observation matrix C are developed, and the effect that these different choices have on the results is examined. Finally, Experiment 3 (Section 3.4) looks at how the method may be modified if observations are not available at every timestep.

3.1 The theta method for the 1D heat equation

The 1D heat equation on $x \in [0,1]$ with a point heat source of strength $\frac{1}{3}$ at $x = \frac{1}{4}$ is,

$$w_t = \sigma w_{xx} + \frac{1}{3}\delta(x - \frac{1}{4}), \tag{3.1}$$

where δ is the Dirac delta function. For this equation, with initial and boundary conditions

$$w(x,0) = f(x), \qquad w(0,t) = w_a, \quad w(1,t) = w_b, \tag{3.2}$$

the familiar theta method discretisation is [8]

$$w_j^{n+1} - w_j^n = \frac{\sigma\Delta t}{\Delta x^2}\left\{(1-\theta)\delta^2 w_j^n + \theta\delta^2 w_j^{n+1}\right\} + s_j\Delta t \qquad 0 \le \theta \le 1 \tag{3.3}$$

with initial and boundary conditions

$$w_j^0 = f(j\Delta x), \qquad w_0^n = w_a, \quad w_J^n = w_b, \tag{3.4}$$

where

$$w_j^n \approx w(j\Delta x, n\Delta t); \quad j = 0,1,..,J = \frac{1}{\Delta x}, \quad n = 0,1,..,N = \frac{1}{\Delta t},$$

$\delta^2 w_j^n$ denotes $w_{j-1}^n - 2w_j^n + w_{j+1}^n$, and s_j is the j^{th} component of the vector $\mathbf{s}$ which is the approximation to the source term.

Discretisation of the source term

We want the approximation to the source term

$$s(x) = \frac{1}{3}\delta(x - \frac{1}{4}), \qquad x \in [0,1] \tag{3.5}$$

to imitate the following two features of $s(x)$;

1.

$$s(x) \begin{cases} > 0 & \text{if} \quad x = \frac{1}{4} \\ = 0 & \text{if} \quad x \neq \frac{1}{4} \end{cases} \tag{3.6}$$

2.

$$\int_0^1 s(u)du = \frac{1}{3}. \tag{3.7}$$

If we choose the vector $\mathbf{s}$ so that its j^{th} component s_j is given by

$$s_j = \begin{cases} \frac{1}{3\Delta x} & \text{if} \quad j = \frac{J}{4} \\ 0 & \text{otherwise,} \end{cases} \tag{3.8}$$

then s_j is a good approximation to $s(j\Delta x)$ in (3.6) as $\Delta x \to 0$. Note also that

$$\sum_{j=1}^{J} s_j \Delta x = \frac{1}{3}, \tag{3.9}$$

where the left hand side is the rectangular rule approximation to the left hand side of (3.7), given that $s_j \approx s(j\Delta x)$ for Δx small. Hence in the limit as $\Delta x \to 0$, (3.7) is satisfied.

The discrete model

The discretisation can be written in matrix form as follows;

$$E\mathbf{w}^{n+1} = A\mathbf{w}^n + \mathbf{u} \tag{3.10}$$

where

$$\mathbf{w}^n = (w_1^n, w_2^n, ..., w_{J-1}^n)^T, \tag{3.11}$$

and where the vector $\mathbf{u}$ containing the boundary conditions and the source term has the form

$$\mathbf{u} = (w_a, 0, ..., \frac{\Delta t}{3\Delta x}, 0, ..., w_b)^T, \tag{3.12}$$

(the non-zero elements of $\mathbf{u}$ are u_j where $j = 1$, $j = J/4$ and $j = J-1$). The $(J-1) \times (J-1)$ tridiagonal matrices E and A are given by

$$E = \begin{pmatrix} (1+2\mu\theta) & -\mu\theta & & & \\ -\mu\theta & (1+2\mu\theta) & -\mu\theta & & \\ & \vdots & \vdots & \vdots & \\ & & \vdots & \vdots & \vdots \end{pmatrix}, \tag{3.13}$$

$$A = \begin{pmatrix} (1-2\mu(1-\theta)) & \mu(1-\theta) & & & \\ \mu(1-\theta) & (1-2\mu(1-\theta)) & \mu(1-\theta) & & \\ & \vdots & \vdots & \vdots & \\ & & \vdots & \vdots & \vdots \end{pmatrix}, \qquad (3.14)$$

where $\mu = \sigma\frac{\Delta t}{\Delta x^2}$.

The theta method is stable $\forall\mu$ if $\frac{1}{2} \leq \theta \leq 1$ and for $0 \leq \mu \leq \frac{1}{2(1-2\theta)}$ if $0 \leq \theta < \frac{1}{2}$, see [8].

In the experiments described here, the model (3.10) is implemented using

$$\Delta x = \frac{1}{16}, \quad \Delta t = \frac{1}{80}, \quad \sigma = 0.1, \qquad \text{so that} \quad \mu = 0.32. \qquad (3.15)$$

The boundary conditions w_a and w_b are taken to be zero. The "true" solution used for generating the observations is obtained by starting the model with initial conditions

$$w_j^0 = 2 \qquad j = 1, .., J-1. \qquad (3.16)$$

The observer, however, is initiated with

$$w_j^0 = 1 \qquad j = 1, .., J-1 \qquad (3.17)$$

at the same time $t = 0$. The observer is expected to converge in time to the observations. For comparison, the solution of the model equations initiated with the "wrong" initial conditions and no observer is also computed.

The dimension n of the system is $J - 1 = 15$. For technical reasons, the observer algorithm as implemented here does not work for n observations. The number of observations p is therefore taken in the range 1 to 14 in these experiments, and different values of $\theta \in [0, 1]$ are tested. Plots were produced comparing the observer solution (solid line) to the numerical solution with no observer (dashed line) and the true solution (plotted ooo). These solutions were plotted against x and shown for $t = 0.25$, $t = 0.5$, $= 0, 75$, and $t = 1$; ie, at 20, 40, 60 and 80 timesteps. The figures referred to here in the text can be found at the end of the report.

3.2 Experiment 1: Changing the eigenvalues

As discussed in Section 2.1 the observer (2.3) we wish to construct will converge to the "true" solution provided that the eigenvalues of $E^{-1}(A - GC)$ have modulus less than unity. Apart from this restriction, we are free to choose this set Λ of eigenvalues as we like. The aim of Experiment 1 is to try out different choices of eigenvalues for the set Λ, and to see which choice gives the fastest convergence of the observer solution to the observations.

The choice of the observation matrix C used in this experiment is rather arbitrary; if there are p observations, then C is the $p \times 15$ matrix

$$C = \begin{pmatrix} 1 & 0 & 0 & \dots & & \\ 1 & 1 & 0 & 0 & \dots & \\ 1 & 1 & 1 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & & & \end{pmatrix}. \tag{3.18}$$

In Experiment 2, however, different observation matrices are developed.

The best choice of eigenvalues investigated in this case was the system eigenvalues (the eigenvalues of $E^{-1}A$) reduced in modulus by 0.25. The system eigenvalues for various values of θ are listed in Table 1 of the Appendix. These results were very pleasing; the observer solution always converged to the true solution by 60 timesteps, so far as could be judged from looking at the plots. Even at 20 timesteps, the solution was generally quite close to the observations. The results were good for all values of θ and for all values of p, indicating that with this choice of eigenvalues, the method works well even if just one or two observations are used. Figure 1 illustrates the quick convergence in the case $\theta = 0$ when 5 observations were used. Since this choice of eigenvalues worked well, it was also used in Experiments 2 and 3.

When the set Λ was chosen to be the system eigenvalues unreduced in modulus, then reasonable results were attained for larger values of p, but the observer solution matched the true solution in fewer and fewer places as p decreased. Other, more arbitrary sets of eigenvalues were also investigated; but these results were less satisfactory. If the eigenvalues were chosen evenly distributed between -0.5 and 0 or between -0.5 and 0.5, then the observer solution only converged within 80 timesteps if enough observations were used; typically 8 or more were needed. Interestingly, these results were poorer than when the system eigenval-

ues were reduced in modulus by 0.25, since both sets of eigenvalues overall had similar modulus. As would be expected, choosing large eigenvalues distributed between -1 and -0.5 gave worse results still; some 11 or more observations were needed for reasonable convergence in this case.

3.3 Experiment 2: Changing the observation matrix

The matrix C can be considered as an interpolation of the model values $\mathbf{w}^k$ from the grid points to the observation points. Choosing the matrix C therefore determines what linear combination of grid point values should be used as the model equivalents to each observation. The theory demands that C should be full row rank (ie, rank p) for constructing the observer. The matrix C used in Experiment 1 was chosen arbitrarily rather than using physical considerations. In Experiment 2, it is supposed that observations are available at anything from 1 to 14 observation points, which do not in general coincide with grid points. The aim of Experiment 2 is to develop an observation matrix which will represent a linear interpolation from the grid point positions to the observation positions. For this, it is at first supposed that the 14 observation positions on the interval $[0, 1]$ were as shown in Table 2 of the Appendix. The matrix C is then built up as follows: if observation i (where $1 \leq i \leq p$) has the position obs_i which lies between grid points x_j and x_{j+1}, then

$$\begin{array}{rcll} C_{i,j} & = & \frac{x_{j+1}-obs_i}{x_{j+1}-x_j}, & \\ C_{i,j+1} & = & \frac{obs_i-x_j}{x_{j+1}-x_j}, & \\ C_{i,k} & = & 0 \qquad k \neq j, \quad k \neq j+1. & \end{array} \tag{3.19}$$

If obs_i lies between either the first or n^{th} grid point and its adjacent boundary point, then row i of C has just one non-zero entry, since the boundary conditions are zero.

For example, with $p = 5$ C is the 5×15 matrix

$$C = \begin{pmatrix} 0.48 & 0 & 0 & ... & & & & \\ 0.08 & 0.92 & 0 & 0 & ... & & & \\ 0 & 0 & 0.96 & 0.04 & 0 & ... & & \\ 0 & 0 & 0 & 0.84 & 0.16 & 0 & ... & \\ 0 & 0 & 0 & 0 & 0.08 & 0.92 & 0 & ... \end{pmatrix}. \tag{3.20}$$

Figure 2 illustrates the performance of the observer in the case $p = 5$, $\theta = 0$. The positions of the observations are marked with a + on the x-axis. Comparing Figures 2 and 1 shows that convergence to the observations is almost as fast as when the matrix C of Experiment 1 is used, in which the observation positions coincide with the grid point positions. This was the case for all values of θ tested, and for $p \leq 12$. If more than 12 observations were used, however, C was no longer of rank p, and the method failed.

Table 3 in the Appendix gives a different set of the observation positions. The same linear interpolation is used, but C no longer has the neat structure of (3.20) since the observations are ordered at random. Figure 3 shows the results obtained in this case with $p = 5$ and $\theta = 0$. The convergence to the observations is much slower than in Figures 1 and 2. The selection of observations at different positions is probably the cause of this. In Figure 2 the observation positions are clustered around the source at $x = \frac{1}{4}$, but in Figure 3 there are no observations near this point. Again, the convergence speeds up when more observations are used, but for $p \geq 12$ the rows of C were linearly dependent, and so the method failed. From the different sets of observation positions tried out, it appears that this happens when there is a cluster of observations between two or three grid points so that one of the rows is a linear combination of the others. This would be less likely to happen if the C matrix had more than just one or two non-zero entries per row.

The results of Experiment 2 show how the observer can successfully deal with observations which do not coincide with the grid points, although the design of the matrix C needs improving so that it maintains full row rank even when p is large. Perhaps using quadratic or higher order interpolation would help here as then each row would have more than two non-zero entries. These results have also shown that two or three observations situated near the source point give much faster convergence than five or more observations away from it. This shows that a couple of carefully placed observations can be more important than a large number of observations randomly situated.

3.4 Experiment 3: Less frequent observations

We now consider the situation where observations are not available at every timestep. This is an important consideration in the context of data assimilation where in practice there will not in general be a complete set of data at every model timestep. Experiment 3a examines the behaviour of the observer after a supply of frequent observations runs out. In Experiments 3b and 3c it is supposed that observations are available every second, fourth or eighth timestep, and two modifications to the data assimilation scheme are considered for dealing with this. The observation positions are as given in Table 2 of the Appendix.

3a) Observations available for the first few timesteps only

It is supposed that observations are available for just 10 or 20 timesteps, but that after this they run out. The observer is employed while the observations are available and is then switched off, and the model run is continued without it.

The success of the results here depended on how close the observer solution was to the true solution when the observer was switched off. Where the observations were available for the first 20 timesteps, the solution stayed close to the truth for all values of p, as Figure 4 illustrates in the case $p = 5$, $\theta = 0$. Where there were observations for only 10 timesteps, the observer solution was not so near to the true solution when the observer was switched off. Nevertheless, in many cases it grew nearer to the true solution as time went on, since the heat equation is dissipative and so smooths out small disturbances. Figure 5 gives an example of a result in which using the observer for just 10 timesteps gave a significantly better solution than if the observer had not been used in the case when $p = 5$, $\theta = 0$. However, in some cases the observer solution was not close to the truth when the observer was switched off, and these results were poorer, this happened when $p = 5$ and $\theta = 0.5$, as Figure 6 shows.

3b) Increasing the timestep

If observations are available at less regular intervals than the timestepping interval, then in some cases the model timestep could be increased so that observations are available at every timestep. This can only be done if the model stability criteria are not violated; that is, only for $\frac{1}{2} \leq \theta \leq 1$ if $\mu = \frac{\sigma \Delta t}{\Delta x^2} > \frac{1}{2(1-2\theta)}$ in our

model.

In this case, the observations are generated from a model run with one value of μ, and the numerical model uses another. So, the observer model not only starts with the wrong initial conditions, but also contains in itself model error, due to the incorrect values of μ. The observer can correct for the wrong initial conditions by driving the model solution to the true solution, but when the observer is switched off, the model solution is expected to drift away from the true solution.

As usual, plots were done for $t = 0.25$, $t = 0.5$, $t = 0.75$ and $t = 1$. However, if observations were available every other timestep, so that the timestep length was doubled, then these plots showed the solutions at 10, 20 , 30 and 40 timesteps. If the observations were available every fourth timestep, the plots showed the solutions at 5, 10, 15 and 20 timesteps. Experiments 1 and 2 show that the observer needs enough timesteps to "settle", so the results at 10 timesteps were not expected to be very good.

Even so, when observations were available every other timestep the results were pleasing, (see Figure 7); on the whole the observer solution converged quite quickly to the true solutions. Generally, the results improved as p increased. When observations were available only every fourth timestep, the results were significantly poorer, (see Figure 8), and if the observations were available every eighth timestep, then satisfactory results were found only for $p \geq 10$.

This approach to the problem of infrequent data is rather unsatisfactory because of its limited practical application, since models generally run with the largest timestep feasible anyway. It also has the disadvantage that p needs to be large when the observations occur infrequently. However, the results from this experiment are interesting in another way. Since the true solution and model have differing values of μ, the model could be seen as having some known error. Some analysis could be done on these results to investigate the effects of model error in data assimilation using ovservers.

3c) Switching the observer on only when observations are available

Here it is supposed that observations are available at regular intervals, and that the observer is switched on when observations were available, and off otherwise. This means using matrix $A_c = (A - GC)$ in place of A whenever observations become available.

This was tried for observations available every other timestep and every fourth timestep. For some values of p the observer solution quickly converged to the observations, and for others, the observer solution even diverged. At first there seemed to be no pattern to this, but the following analysis gives an explanation. The matrix G was chosen to assign an eigenstructure so that the error $\mathbf{e}^k$ given by the equation

$$\mathbf{e}^k = E^{-1}(A - GC)\mathbf{e}^{k-1} = (E^{-1}(A - GC))^k \mathbf{e}^0 \tag{3.21}$$

would tend to zero as $k \to \infty$. However, if the model alternates between A_c and A at successive timesteps, then the error equation becomes

$$\mathbf{e}^{2k} = \{E^{-1}AE^{-1}(A - GC)\}^k \mathbf{e}^0. \tag{3.22}$$

Hence in this case the observer matrix G should be chosen so that the eigenvalues of $E^{-1}AE^{-1}(A - GC)$ all have modulus less than one. It was found that in the cases where the observer solution converged towards the truth, this condition held, and where it diverged, at least one eigenvalue was greater than one.

These results indicate that with the modifications suggested above, the observer data assimilation scheme could work well with observations at less frequent intervals.

4 Data assimilation using successive correction

Some of the earliest attempts at data assimilation in the late 1950s used an approach known as "successive correction". Since then, this conceptionally simple approach has been developed into schemes which are sometimes quite sophisticated. The basic idea is to modify the model solution in the light of the observations. In its simplest form, this means adding some proportion of the difference between an observation and its model counterpart to the model solution at all grid points within some "radius of influence" of the observation. In the Cressman scheme [2], the proportion to be added to a particular grid point depends on its distance to the observation. If w_{ij} is the weight or proportion of the correction to grid point i with respect to observation point j, then

$$w_{ij} = \begin{cases} \frac{1}{\rho_i}\frac{R-d_{ij}}{R+d_{ij}} & \text{if} \quad d_{ij} < R \\ 0 & \text{if} \quad d_{ij} \geq R \end{cases} \tag{4.1}$$

Here R is the radius of influence, and d_{ij} is the distance of grid point i to observation j, and ρ_i is the number of nonzero entries in row i. In [2] the correction stage of each model timestep is repeated several times with successively smaller values of R. Here just one value of R is used since this is only a small scale problem.

The weight w_{ij} forms the ij^{th} element of the Cressman weighting matrix W. This section discusses how W relates to the observer feedback matrix G, and compares the performance of the Cressman scheme to the observer for the same model as used in Section 3.

4.1 Comparison of the observer and successive correction techniques

Suppose the true state of the atmosphere is described by the discrete linear time invariant system

$$\mathcal{S}: \qquad E\mathbf{w}^{k+1} = A\mathbf{w}^k + B\mathbf{u}^k \tag{4.2}$$

and that we have observations $\mathbf{y}^k$ of the state $\mathbf{w}^k$ given by

$$\mathbf{y}^k = C\mathbf{w}^k. \tag{4.3}$$

In a general sucessive correction method, each model timestep involves two stages: a model update, and then a correction. Writing $\tilde{\mathbf{w}}^{k+1}$ for the updated model state,

and $\hat{\mathbf{w}}^{k+1}$ for the corrected model state, we have

Stage 1: model update

$$E\tilde{\mathbf{w}}^{k+1} = A\hat{\mathbf{w}}^k + B\mathbf{u}^k, \tag{4.4}$$

Stage 2: correction

$$\hat{\mathbf{w}}^{k+1} = \tilde{\mathbf{w}}^{k+1} + W(\mathbf{y}^{k+1} - C\tilde{\mathbf{w}}^{k+1}). \tag{4.5}$$

Substituting (4.2) - (4.4) into (4.5) gives

$$\hat{\mathbf{w}}^{k+1} = E^{-1}(A\hat{\mathbf{w}}^k + B\mathbf{u}^k) + WCE^{-1}(A\mathbf{w}^k + B\mathbf{u}^k - A\hat{\mathbf{w}}^k - B\mathbf{u}^k), \tag{4.6}$$

and subtracting (4.6) from (4.2) gives

$$\mathbf{w}^{k+1} - \hat{\mathbf{w}}^{k+1} = E^{-1}A(\mathbf{w}^k - \hat{\mathbf{w}}^k) - WCE^{-1}A(\mathbf{w}^k - \hat{\mathbf{w}}^k). \tag{4.7}$$

Defining $\mathbf{e}^k = \mathbf{w}^k - \hat{\mathbf{w}}^k$; we have the error equation

$$\mathbf{e}^{k+1} = (I - WC)E^{-1}A\mathbf{e}^k. \tag{4.8}$$

From this, we can see that a necessary condition for the state to converge to the observations is that $(I - WC)E^{-1}A$ should have eigenvalues of modulus less than unity. The corresponding error equation obtained in Section 2, which gives conditions for the choice of G, is

$$\mathbf{e}^{k+1} = E^{-1}(A - GC)\mathbf{e}^k. \tag{4.9}$$

Comparing the two error equations (4.8) and (4.9) gives a way of comparing the Cressman scheme with the taditional control theory observer. The main difference between the two approaches is that W acts on $\mathbf{y}^{k+1}$ and $C\mathbf{w}^{k+1}$, but G acts on $\mathbf{y}^k$ and $C\mathbf{w}^k$. This explains why $E^{-1}GC$ in (4.9) is replaced by $WCE^{-1}A$ in (4.8), since $E^{-1}A$ represents a model update from $\mathbf{w}^k$ to $\mathbf{w}^{k+1}$. Writing $\tilde{C}$ in place of $CE^{-1}A$, (4.8) can be written

$$\mathbf{e}^{k+1} = E^{-1}(A - EW\tilde{C})\mathbf{e}^k. \tag{4.10}$$

Comparing (4.10) with (4.9) and considering $\tilde{C}$ as a modified observation matrix, it can be seen that the role of the feedback matrix G is taken by EW in the Cressman scheme.

4.2 Experiments with the Cressman scheme

The Cressman scheme as described in (4.1) was implemented for the system (4.2) to compare its performance with the observer G developed in Section 2. In this experiment, the matrices E, A, B, C; the input $\mathbf{u}$ and the initial condition $\mathbf{w^0}$ were chosen as in Section 3.1. The Cressman scheme was tested for different values of θ and p, and for different values of R, the radius of influence. As before, plots were produced comparing the "Cressman solution" (solid line) with the original numerical solution (dashed line) and the "true solution" (plotted ooo), at 20, 40, 60, and 80 timesteps.

For all values of θ, the success of the scheme for driving the model solution to the true solution depended strongly on the number of observations used. For $R = 0.3$, using just one observation had almost no impact on the solution, and using 3 observations gave some improvement to the numerical solution (see Figure 9). If 5 or more observations were used, the solution was quite near to the truth at 80 timesteps (Figure 10), and the more observations that were used, the sooner the solution converged to the truth (Figure 11). Increasing the radius of influence also improved the results, although not as much as increasing the number of observations used.

When p and R were very small, the eigenvalues of $(I - WC)E^{-1}A$ were the same as the system eignvalues, that is the eigenvalues of $E^{-1}A$. Increasing p decreased the largest few eigenvalues, but the smallest were left unchanged. This corresponds to the significant improvement of the results as p increases. Increasing the radius of influence gave a very small decrease in the largest eigenvalue. However, since the error at timestep k depends on the eigenvalues raised to the power k, the small difference in the largest eigenvalues becomes significant when the eigenvalues are raised to the power of 20 or 40. This corresponds to the improvement seen in the results as R increases.

5 Conclusions and suggestions for further work

A robust observer has been successfully built for a very simple model in the context of data assimilation. It was found that a good eigenvalue assignment for the system with the observer is given by reducing the modulus of the eigenvalues of the original model system by 0.25. This choice gave fast convergence of the observer to the true solution in the cases investigated, even if only one or two observations were used.

The observation matrix C developed to represent a linear interpolation from the model states to the observation positions worked well, but if too many observations were clustered around two or three grid points, then the rows of C were no longer linearly independent. This could perhaps be prevented if a higher order interpolation were used. The results of Experiment 2 also point out that it can be more important to have a couple of carefully placed observations than a larger number of observations positioned randomly.

The results from Experiment 3 showed that the observer could also work well if observations are not available at every timestep, which is a very important practical property for operational data assimilation. The method of switching the observer on and off, and choosing suitable eigenvalues so that this would work, is a promising way to cope with this problem, but needs to be developed. Extending the model timestep so that it coincides with the frequency of the observations is not of much practical interest, since generally models already run at the largest timestep possible. However, the results from this experiment show that even though the model contains error because it runs with the wrong value of μ, the observer can still be driven to fit the observations.

Checking that the solution stays close to the truth after the observations have run out is important, since in numerical weather prediction model runs using data assimilation are required to provide initial conditions for a forecast for which there are no observations. If the model contains error, though, the model solution would diverge from the truth after the observer is switched off.

In Section 4, a general successive correction scheme was linked to the observer technique described in Section 2, and a link between the feedback matrix G and the Cressman weighting matrix W was established. Error analysis on the Cressman scheme gave conditions for convergence of the scheme to the observa-

tions. Reformulating data assimilation schemes in terms of observers introduces the possibility of using results from control theory to carry out analysis on these schemes. This could perhaps explain some of the successes and failings of the different schemes. The experiments described using the Cressman scheme showed the robust observer developed in this project to work better than the Cressman scheme when there were few observations. This can be linked to the fact that the error equation for the observer has smaller eigenvalues than that of the Cressman scheme.

A major probem with the observer developed here is that it is based on assigning the eigenstructure of the system, but finding the eigenvalues and eigenvectors of a large numerical model is no trivial matter! Hopefully, though, comparing this method to other data assimilation schemes applied to simple models will give insights for improvements to those methods. There are other methods for developing observers, and [7] develops an observer for a nonlinear model using techniques which appear to bear resemblence to the variational analysis approach to data assimilation, and to the methods we have investigated in another report [5] on optimal control theory and data assimilation. It would be interesting to investigate this approach to building an observer next.

References

[1] S. Barnett and R.G. Cameron. *Introduction to Mathematical Control Theory.* OUP, second edition, 1984.

[2] G.P. Cressman. An operational objective analysis system. *Monthly Weather Review*, pages 367–374, Oct 1959.

[3] L.R. Fletcher, J. Kautsky, and N.K. Nichols. Eigenstructure assignment in descriptor systems. *IEEE Trans Auto Cnt*, AC-31, 1986.

[4] M. Ghill and P. Malalnotte-Rissoli. Data assimilation in meteorology and oceanography. *Adv Geophys*, 33:141–266, 1991.

[5] A.K. Griffith and N.K. Nichols. Data assimilation using optimal control theory. Numerical Analysis Report 10/94, The University of Reading, 1994.

[6] J. Kautsky, N.K. Nichols, P. van Dooren, and L. Fletcher. Numerical methods for robust eigenstructure assignment in control system design. In *Proc. of Workshop on Numerical Treatment of Inverse Problems for Integral and Differential Equations*, Heidelburg, 1982.

[7] H. Michalska. An observer for nonlinear descriptor systems of index one. In *Proceedings of the International Symposium on Implicit and Nonlinear Systems*, pages 372–379, 1992.

[8] G.D. Smith. *Numerical Solution of Partial Differential Equations: Finite Difference Methods.* OUP, third edition, 1985.

[9] S.M. Stringer. *The Use of Robust Observers in the Simulation of Gas Supply Networks.* PhD thesis, The University of Reading, Department of Mathematics, 1993.

APPENDIX

Table 1: System eigenvalues

The eigenvalues of $E^{-1}A$ defined in equation (3.10) for $\theta = 0$, $\theta = 0.5$ and $\theta = 1$ are:

$\theta = 0$	$\theta = 0.5$	$\theta = 1$
0.7156	0.9878	0.9879
0.6045	0.9524	0.9026
0.8125	0.8977	0.9535
0.8921	0.8286	0.8421
0.9513	0.7510	0.7786
0.9877	0.6701	0.7168
0.4849	0.5904	0.6600
0.3600	0.5152	0.6098
0.2351	0.4467	0.5666
0.1151	0.2241	0.6305
0.0044	0.2379	0.5011
−0.0925	0.3353	0.4410
−0.1221	0.2934	0.4604
−0.2313	0.3865	0.4482
−0.2677	0.2610	0.4770

Table 2: First set of observation positions

obs_1	obs_2	obs_3	obs_4	obs_5	obs_6	obs_7	obs_8	obs_9	obs_{10}	obs_{11}	obs_{12}	obs_{13}	obs_{14}
0.03	0.12	0.19	0.26	0.37	0.42	0.45	0.56	0.57	0.60	0.67	0.71	0.73	0.83

Table 3: Second set of observation positions

obs_1	obs_2	obs_3	obs_4	obs_5	obs_6	obs_7	obs_8	obs_9	obs_{10}	obs_{11}	obs_{12}	obs_{13}	obs_{14}
0.70	0.03	0.45	0.92	0.80	0.12	0.15	0.20	0.56	0.90	0.87	0.09	0.63	0.43

List of Figures

Key to the figures:

Observer solution

\- - - Numerical solution with no observer

o o o True solution

\+ + + Positions of the observations

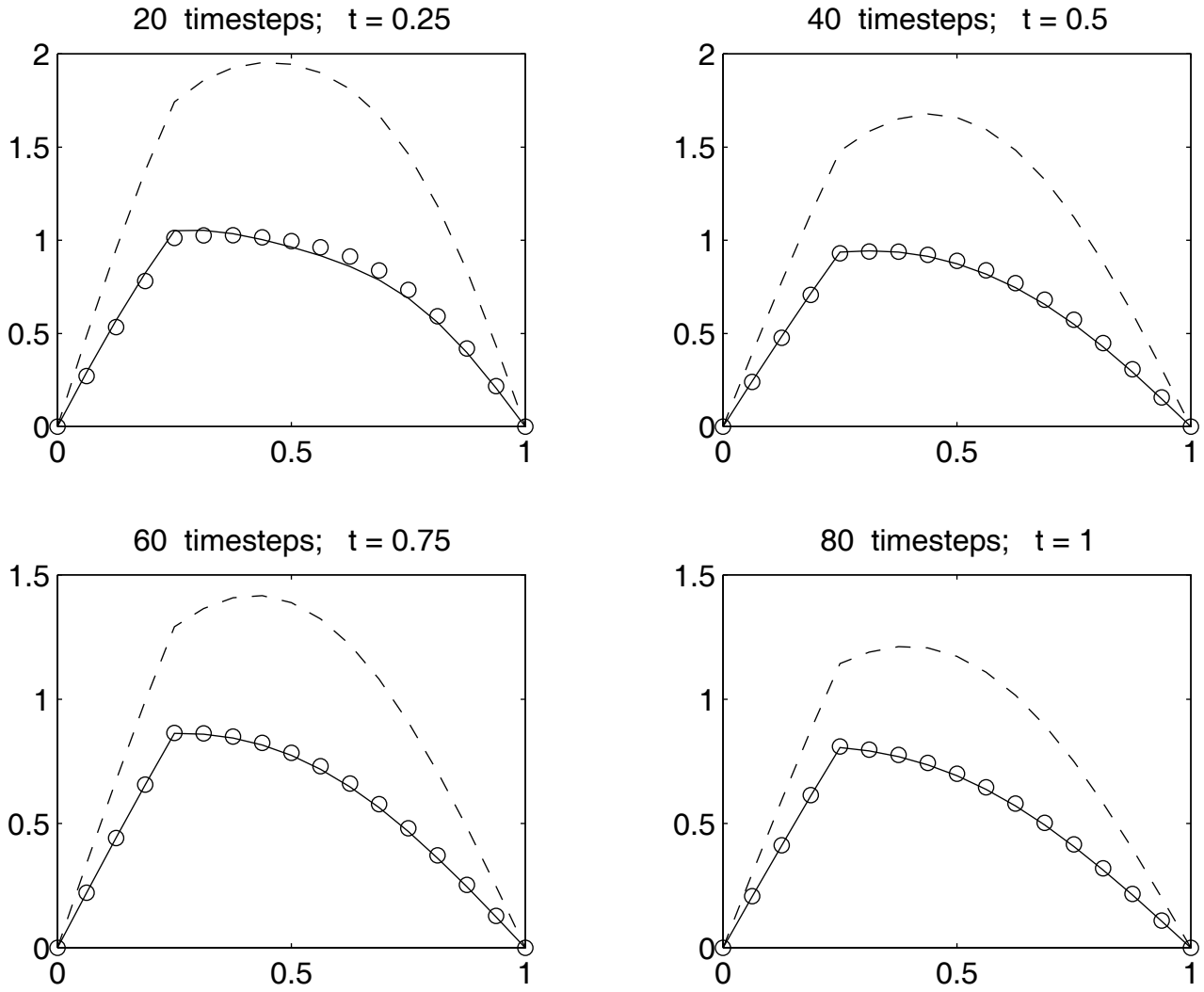

Figure 1: Experiment 1; $p = 5$, $\theta = 0$

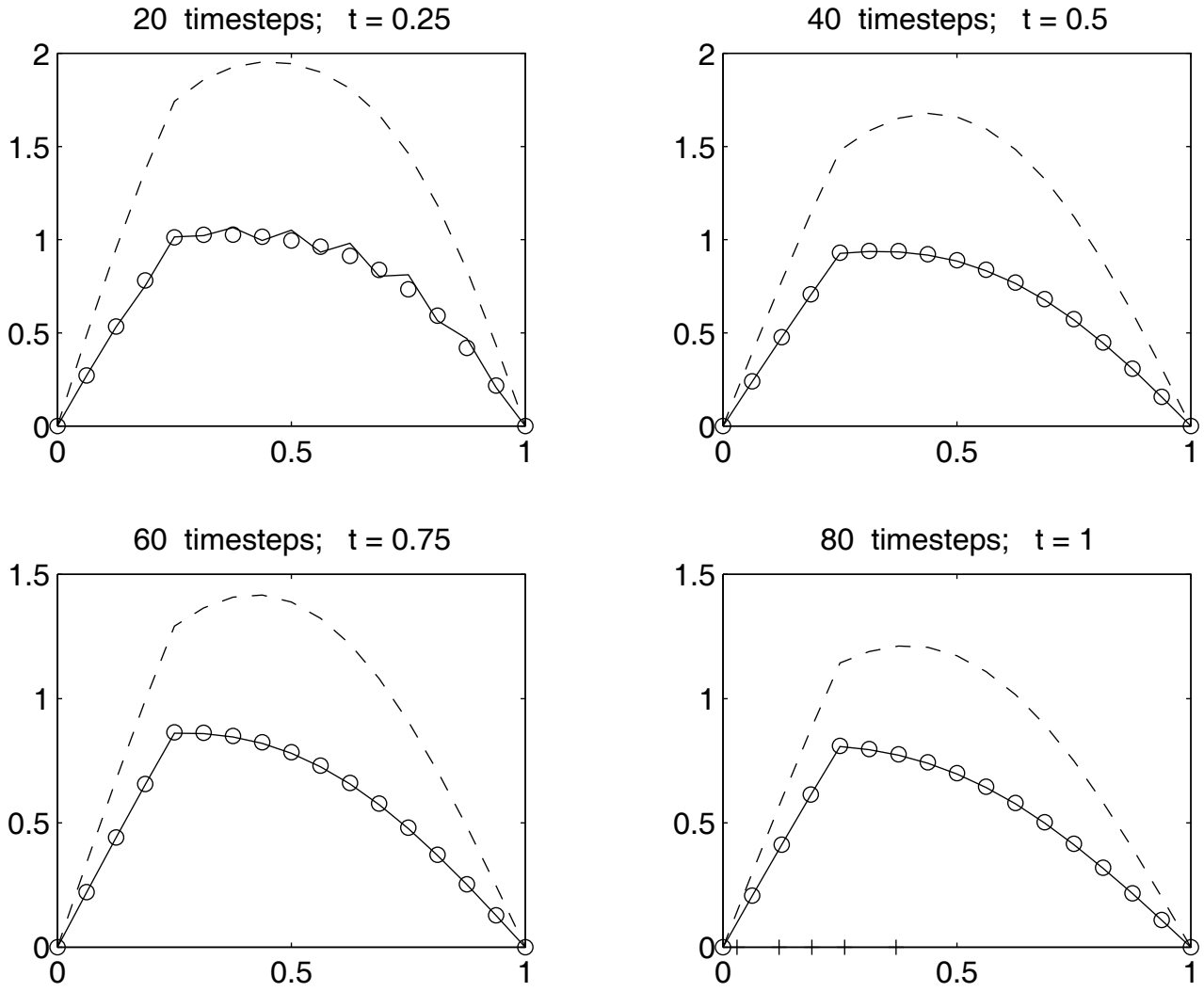

Figure 2: Experiment 2; $p = 5$, $\theta = 0$. First set of observation positions (ordered)

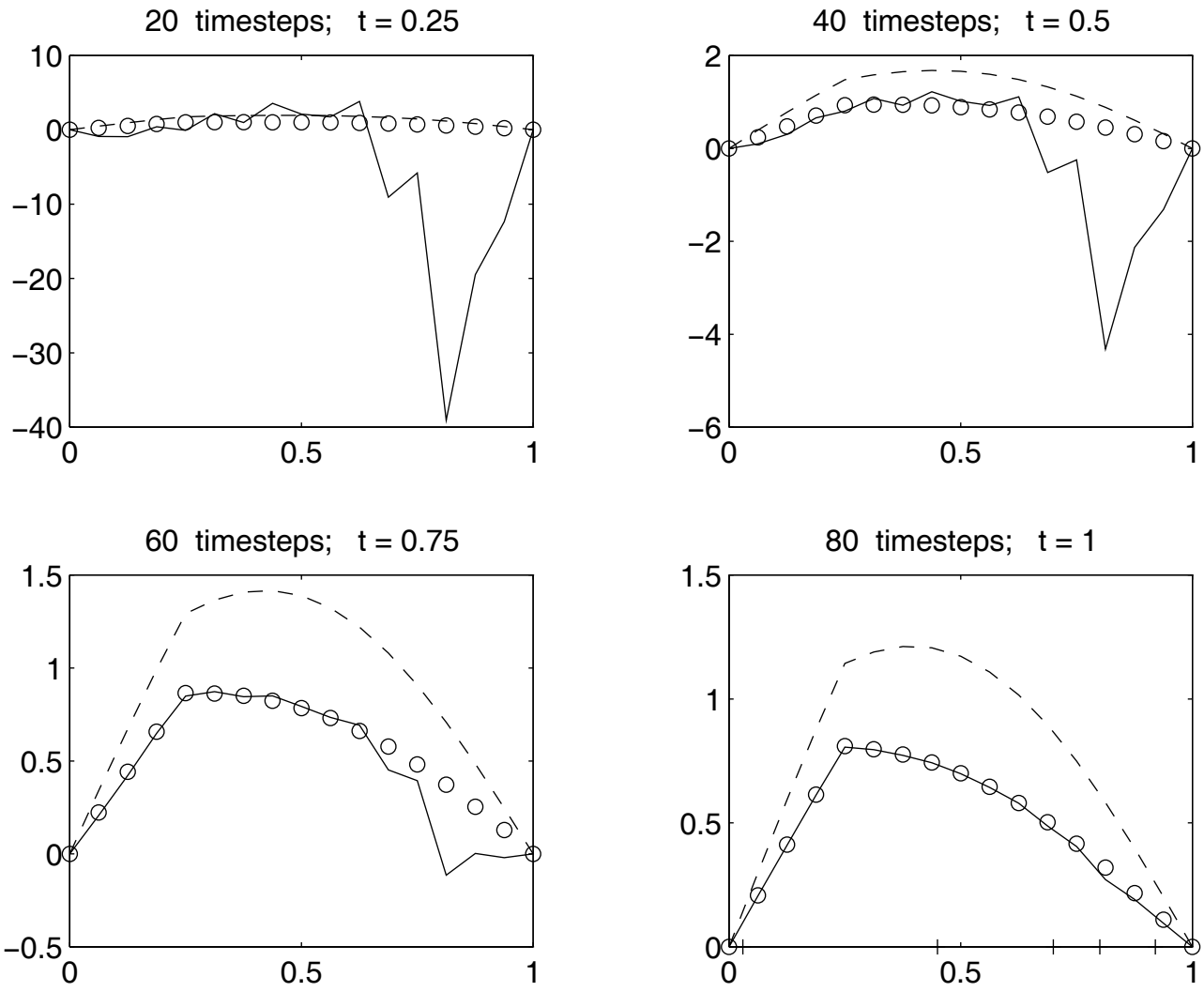

Figure 3: Experiment 2; $p = 5$, $\theta = 0$. Second set of observation positions (unordered)

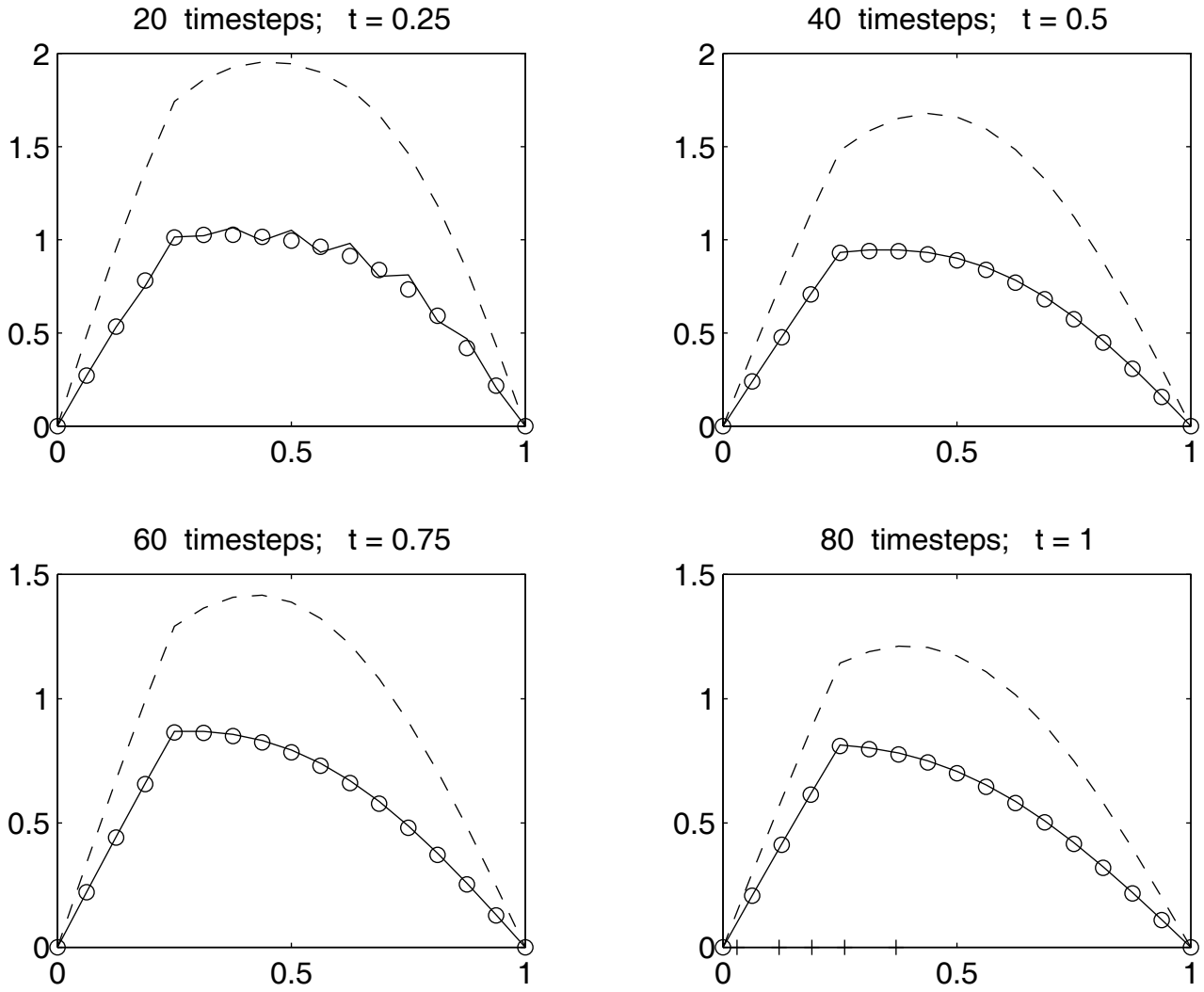

Figure 4: Experiment 3a; $p = 5$, $\theta = 0$. Observations available for the first 20 timesteps

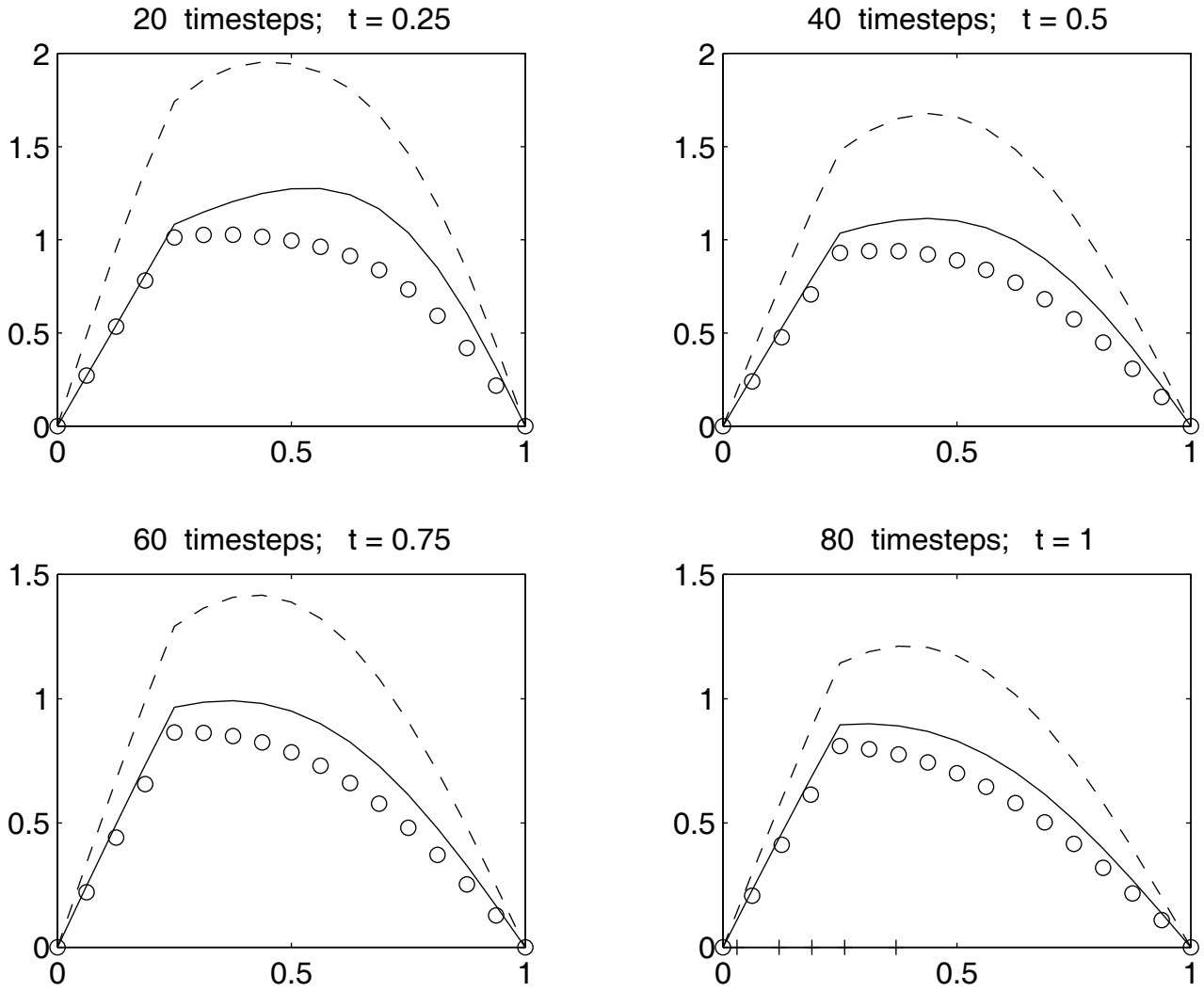

Figure 5: Experiment 3a; $p = 5$, $\theta = 0$. Observations available for the first 10 timesteps

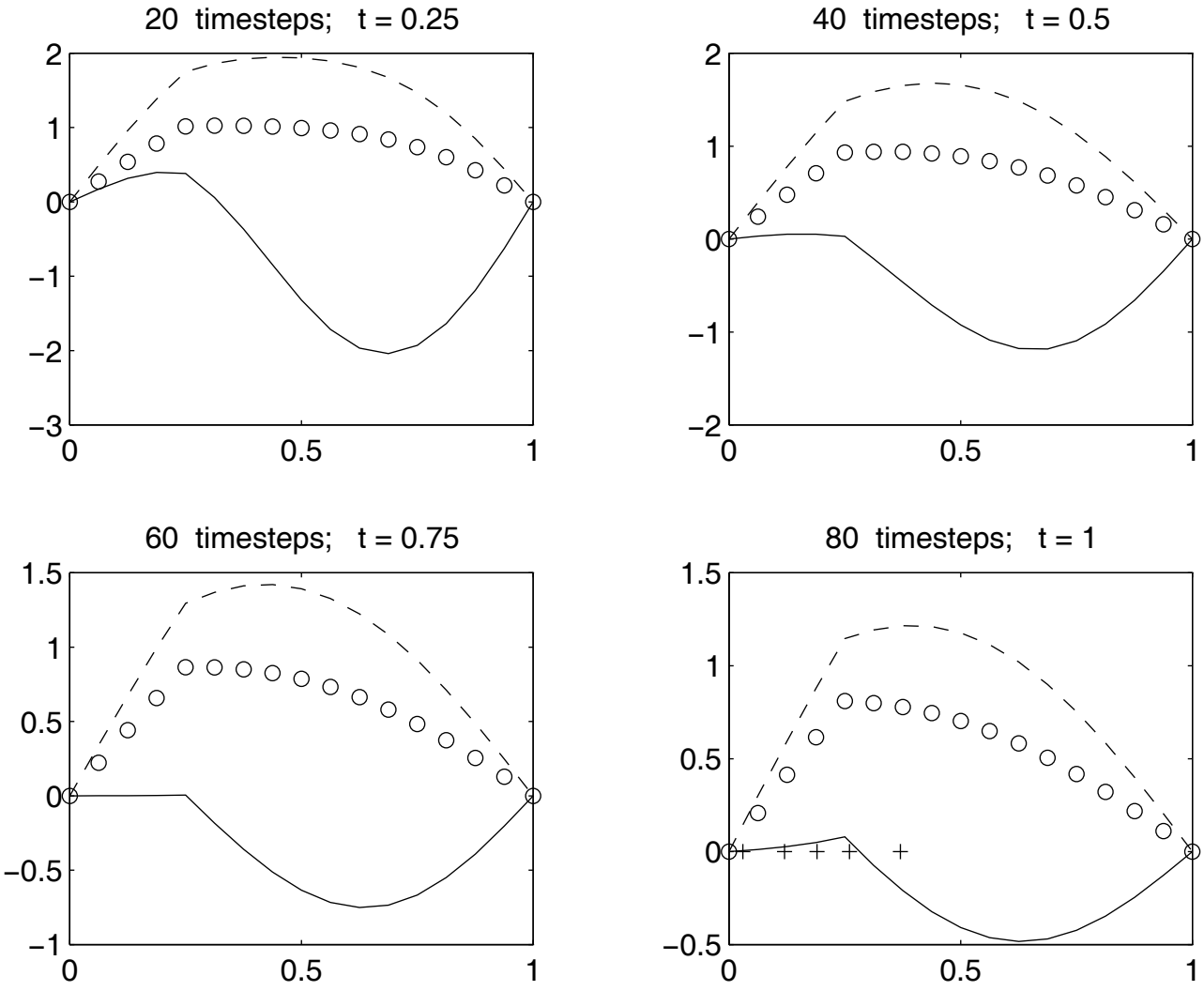

Figure 6: Experiment 3a; $p = 5$, $\theta = 0.5$. Observations available for the first 10 timesteps

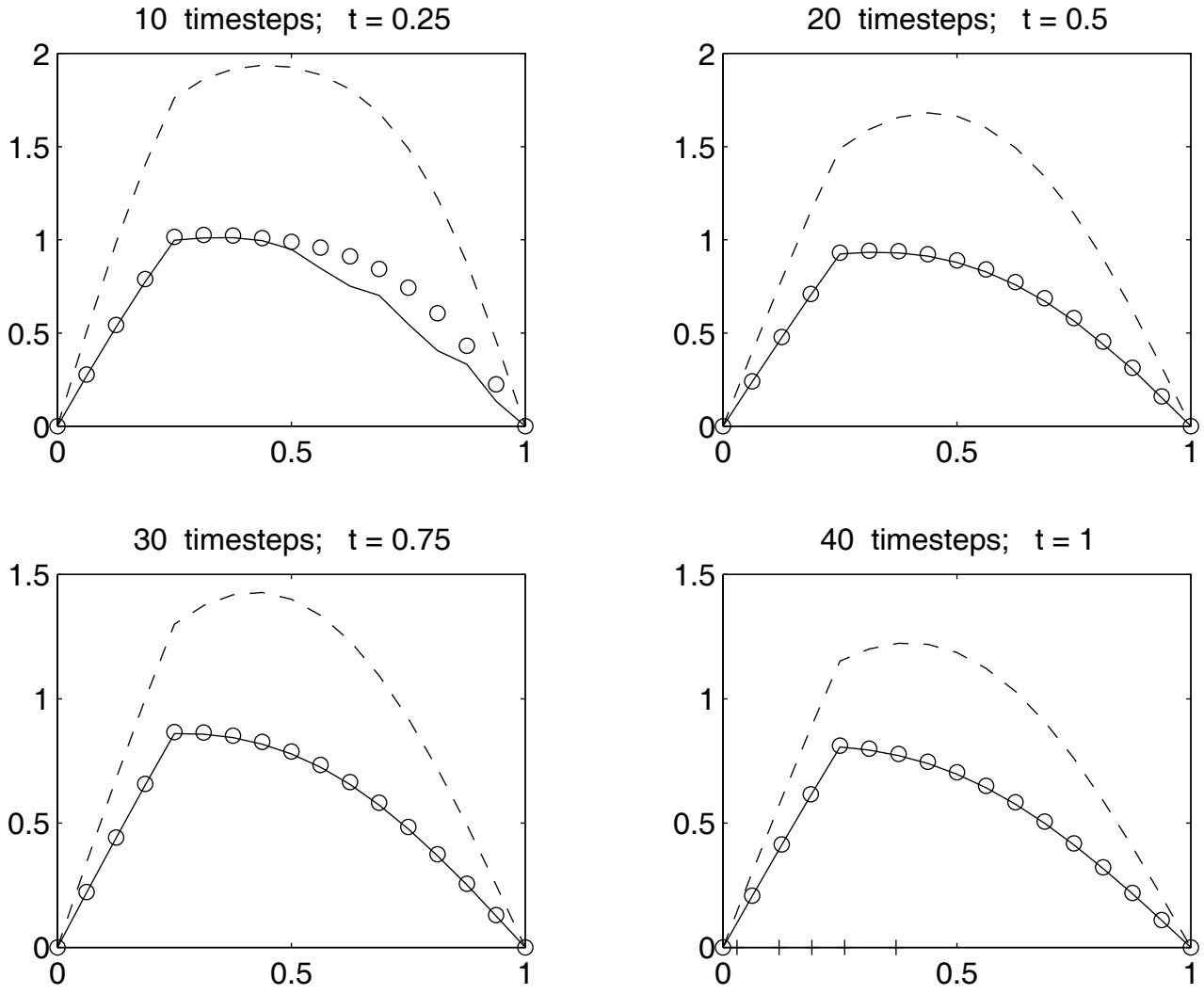

Figure 7: Experiment 3b; $p = 5$, $\theta = 1$. Observations available every second timestep

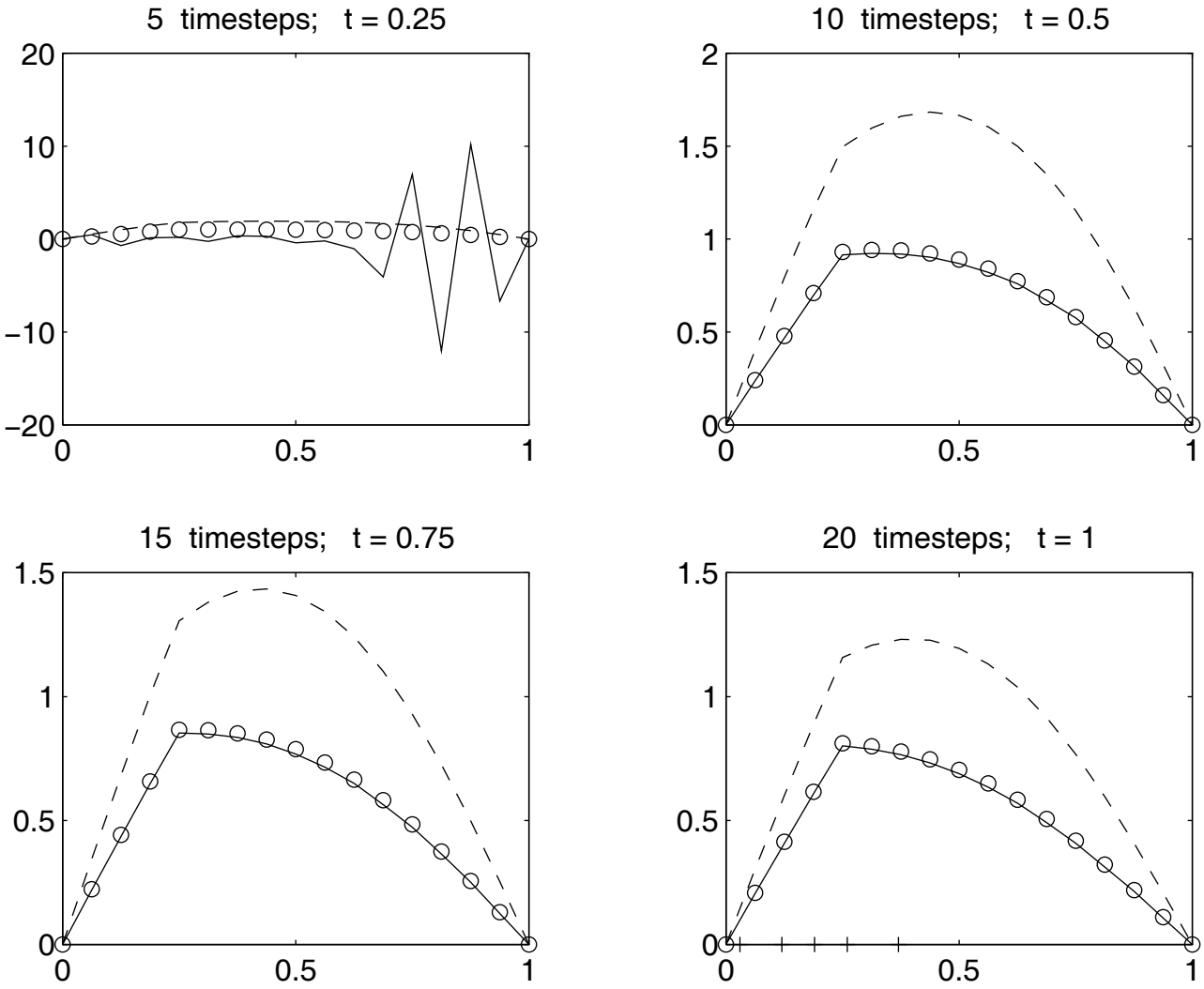

Figure 8: Experiment 3b; $p = 5$, $\theta = 1$. Observations available every fourth timestep

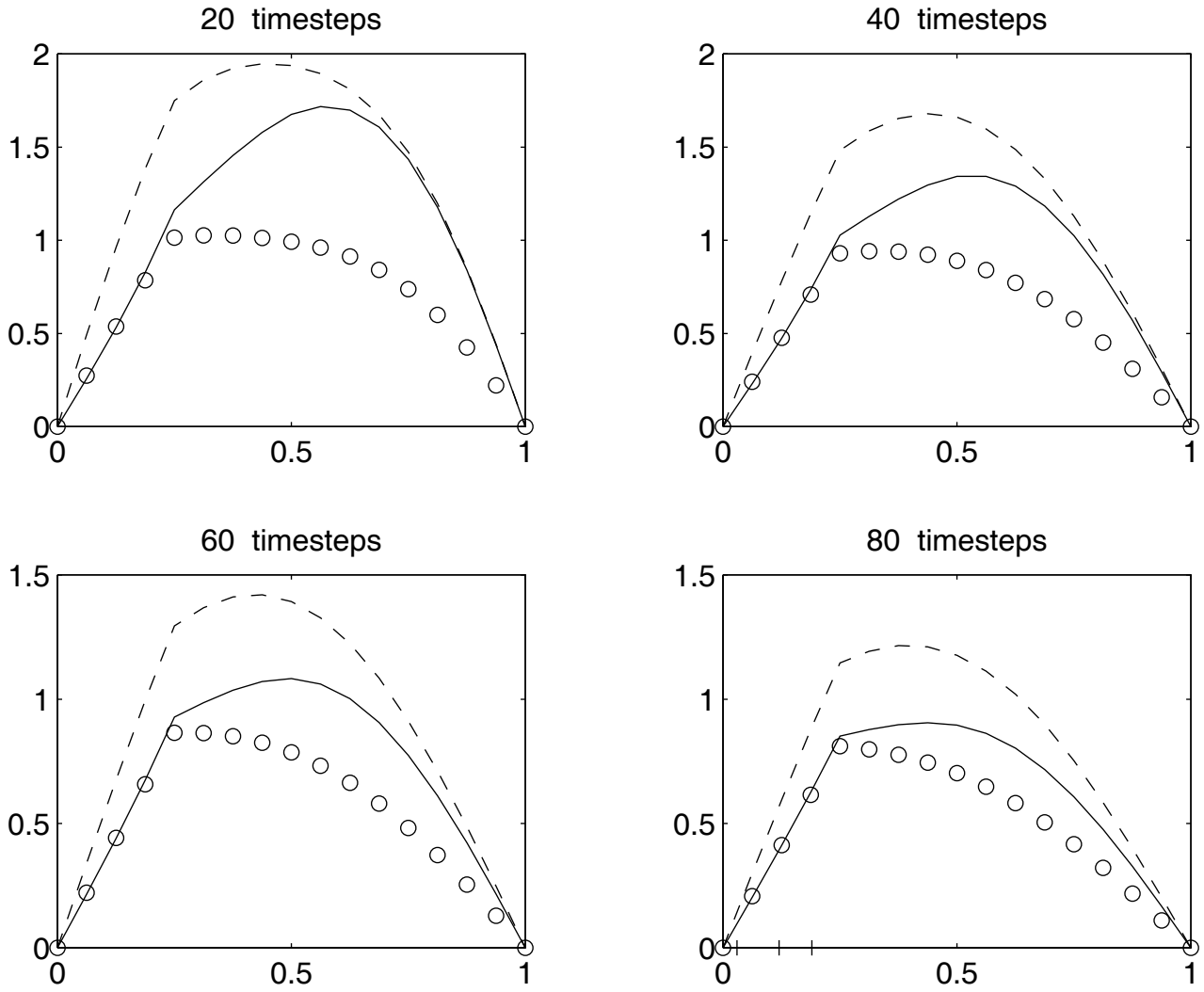

Figure 9: Cressman scheme; $p = 3$, $\theta = 0.5$, $R = 0.3$

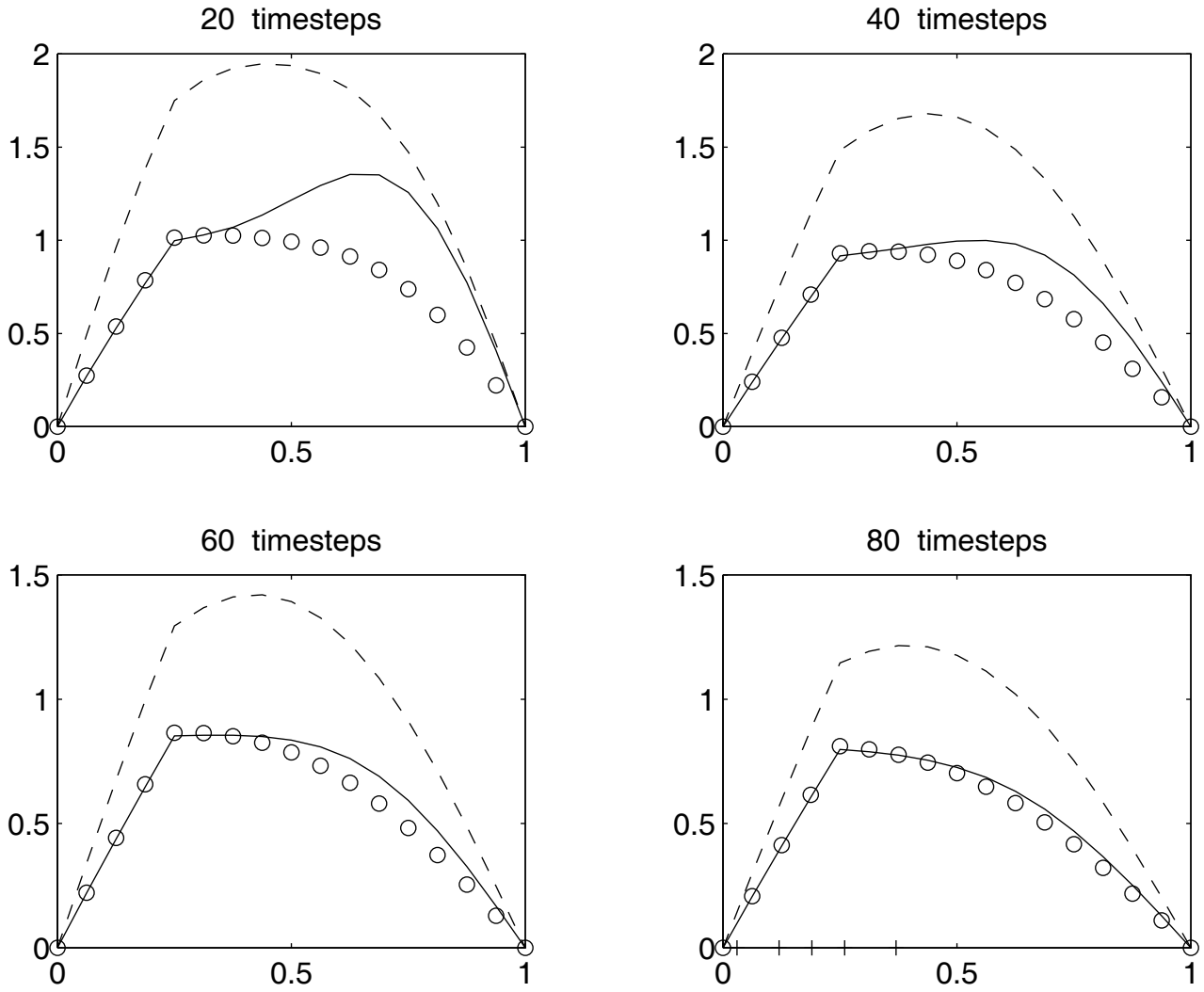

Figure 10: Cressman scheme; $p = 5$, $\theta = 0.5$, $R = 0.3$

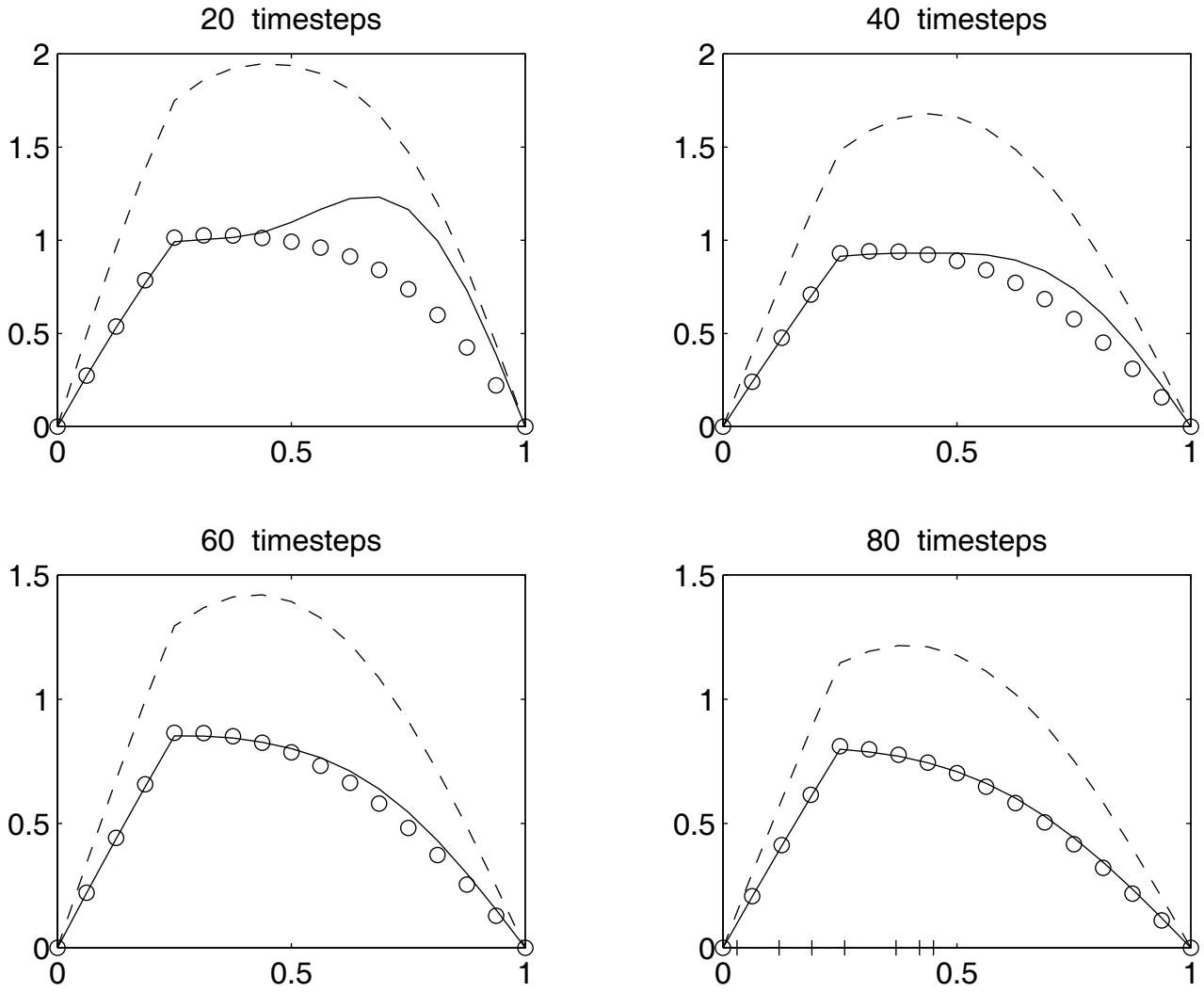

Figure 11: Cressman scheme; $p = 7$, $\theta = 0.5$, $R = 0.3$