OPTIMAL NONLINEAR ANALYSIS

Andrew C Lorenc
Met O 11, Meteorological Office, Bracknell.

SUMMARY: Some practical problems in objective analysis for numerical weather prediction are best solved by using nonlinear analysis equations. They include the existence of nonlinear prior constraints on the analysis, and the use of observations which are nonlinearly related to the analysis parameters, or which have non-Gaussian error distributions. The behavior of nonlinear analysis equations is demonstrated with a simple one-dimensional shallow-water model. It is shown that time-tendency information, and indirect observations such as wind speed, or the movement of a tracer, can be used in the analysis. The resulting forecasts are better than those made from an analysis from a traditional analysis-forecast cycle. The nonlinear method is shown to be capable of "moving" a discontinuity similar to a front, to fit observations defining its position, thus giving an analysis with more detail than would be expected from the spatial resolution of the observations. The incorporation of additional nonlinear constraints, such as that used in initialization, is demonstrated. The method can be used to effectively reject observations with gross errors, by specifying a non-Gaussian error distribution. However this generates multiple minima which complicate the search for the best analysis, so the complex decision taking algorithms associated with other methods of quality control are not avoided.

1. INTRODUCTION

This paper is based on work published in Lorenc (1988a). Theoretical justifications and details of methods are given there, and in Lorenc (1986). The iterative solution methods used are also discussed in Lorenc (1988b). Here we concentrate on meteorological interpretation of results.

Most methods of objective analysis in use for meteorological data are linear in the observed data values. That is, each analysed value $x_a[j]$ can be written as a linear combination of the data values $y_o[i]$; this is usually done in terms of deviations from background values $x_b[j]$ and $y_b[i]$:

$$x_a[j]=x_b[j]+\sum_i W[j,i]\ (y_o[i]-y_b[i]) \qquad (1)$$

(1) is linear if the weights $W[j,i]$ are independent of the actual observed values y_o, but depend only on their positions and accuracies. This is true of least-squares fitting techniques, of the successive-correction method, and of optimal interpolation. Such equations are optimal if the error distributions of both observations and of any background fields are Gaussian, prior constraints and relationships describing the desired analysis are linear, and the relationships between analysed and observed variables are linear (Lorenc 1986). Such assumptions lead to a variational problem with a quadratic penalty function.

Many of the problems currently of interest in meteorological objective analysis for numerical weather prediction (NWP) do not fit these assumptions. These problems include :-

a. Four-dimensional analysis contrained by nonlinear prognostic equations, to use time-tendency information from observations.

b. Use of "indirect" observations, with a nonlinear relationship between observed parameters and those analysed.

c. Analysis of near-discontinuities, such as fronts, which can be regarded as a strong nonlinear coupling between spatial scales in the atmosphere.

d. Multivariate analysis, subject to nonlinear balance constraints.

e. Quality control of observations, and use of data whose error distributions are known to be non-Gaussian.

The non-quadratic variational problem which results is given in section 2. It can be solved iteratively, using the methods described in Lorenc (1988b), although practical meteorological problems are usually so large that approximations have to be

made. In order to study the full nonlinear optimal equation, we apply it to a simple one-dimensional nonlinear shallow-water equation model, described in section 3. The initial conditions used are such that a hydraulic jump develops; this is about the simplest system that simulates some of the features of atmospheric fronts, and with it we can study all the problems (a) to (e) listed above. This is done in section 4, referring each example to analagous practical problems. It is shown that all of the examples can be solved using the optimal nonlinear analysis equation, so that, in principle at least, the practical problems are capable of solution.

2. NONLINEAR ANALYSIS EQUATION

It is shown in Lorenc (1988a) that the optimal analysis is obtained by minimizing a penalty function J given by:

$$J = -\ln(P_{o\mathbf{f}}(\mathcal{K}(\mathbf{x})-\mathbf{y}_o)) - \ln(P_b(\mathbf{x}-\mathbf{x}_b)) \qquad (2)$$

$\mathcal{K}$ is the generalized interpolation operator which relates analysis variables $\mathbf{x}$ to observed variables $\mathbf{y}$. $P_{o\mathbf{f}}$ is the observational and representativeness error distribution function, and P_b is the background error distribution function; if these are Gaussian the equation reduces to the form given in (1) of Lorenc (1988b). This was solved for this work using conjugate-gradient and Newton descent algorithms.

3. EXPERIMENTAL SYSTEM

In order to study the behavior of the nonlinear analysis equations in an ideal situation, we choose a simple system where the prediction model, its adjoint, and the background and observational error distributions can be known exactly. A one-dimensional finite-difference model of the shallow-water equations is used as a strong constraint on the permitted space- and time-values. Thus we can reduce the analysis problem to that of finding the initial conditions for this model which, when forecast with the model equations, best fit the observations and other constraints. Initialized from smooth initial conditions which are known to generate a hydraulic jump, the model is first integrated forward to

generate a "truth". Background and observational error covariances are defined, and a background and observations are calculated by adding pseudo-random numbers consistent with these. Thus all of the results fall into the class of "identical-twin" experiments.

The one-dimensional shallow-water equations with rotation are used in the flux form given by Parrett and Cullen (1984 eqn.8), who showed that these equations are suitable for simulating a hydraulic jump. The initial conditions, taken from their equation 9, are smooth, but for the parameters chosen develop a hydraulic jump.

Cyclic boundary conditions are used in space, with resolution 128, and the model is integrated until (non-dimensional) time T=2.8, which is long enough for the hydraulic jump to develop and propagate nearly across the domain.

All variables were given the same background error auto-correlation, shown in Fig.1. Cross correlations were assumed to be zero. Observational errors were assumed to be random and Gaussian, with variance .0025 of that of the same variable in the background.

4. RESULTS AND METEOROLOGICAL INTERPRETATION.

In this section we present and discuss optimal nonlinear analyses obtained by minimizing (2) for the system described in section 3. Our object is to demonstrate what is theoretically achievable in an ideal situation. Results are discussed in 5 sub-sections, corresponding to the 5 meteorological analysis problems listed in the introduction. The experiments performed are listed in Table.1 and Table.2.

4a. Four-dimensional data assimilation

Four-dimensional data assimilation can be defined to be the use of a four-dimensional distribution of observations, together with a constraint on the resulting four-dimensional analysis that its evolution in time should satisfy known prognostic equations, as embodied in a forecast model (Lorenc 1986). Since forecast models are not perfect, the latter constraint should not be strictly enforced. Practical implementation of such a scheme is severely limited by

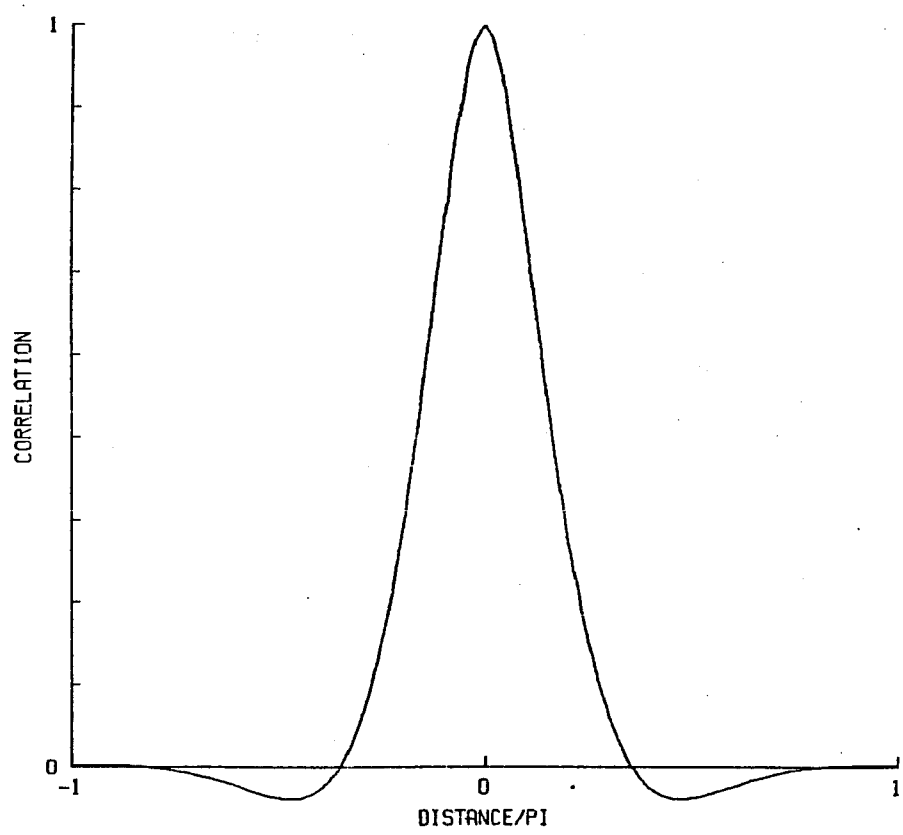

Fig.1. Horizontal correlation of error of background wb used in experiments. Cross-correlations between variables were assumed to be zero. All variables had the same auto-correlation.

available computer resources. The storage and manipulation of high-resolution four-dimensional analyses requires many more resources than running an NWP forecast model, which only manipulates three-dimensional fields.

The traditional NWP method of approximating four-dimensional data assimilation is the analysis-forecast cycle, in which observations are inserted using a purely three-dimensional analysis procedure into a 'background' forecast from the results of the previous analysis. This method does allow for the imperfections of the forecast model; an estimated background error is used to calculate the observation weights in statistical analysis methods such as OI. However it does not properly use time-tendency information in the observations, since observations from different times are not analysed together.

If the prognostic equations are applied as a strong constraint as in section 2c, then some of the computational difficulties can be avoided through a reduction of the control variable. The four-dimensional analysis is defined by a three-dimensional initial field plus the prognostic equations, and (except for the current estimate x_1) four-dimensional fields need not be stored and manipulated, while still obtaining an optimal four-dimensional use of observational information. However this does not allow for inaccuracies in the forecast model over the period of the observations used.

This work combines the latter 'strong constraint' approach for a short time-period, with the traditional analysis-forecast cycle approach, so that time-tendency observations within this period can be used optimally, while information from earlier observations can be used without assuming that the forecast model is perfect. Thus while it is not truly optimal in its application of imperfect prognostic constraints, it does match well the operational NWP requirement of providing initial conditions for a forecast, using the last analysis and observations valid during the period since the last analysis was made.

Fig.2 shows the height field at time T=0 from experiments to illustrate this. In this and most subsequent figures the fields are plotted as solid lines, the "truth" from which

Table 1. Experiments described in sections 4a to 4d.
Curves in Fig.2 to Fig.10 are labelled with Expt.
Observations were either equally distributed in space at each time, or equally distributed in time at one position.
Experiments used a space grid of 128 points.

Expt	Observations	Description
A	35 h at T=0.0 35 h at T=1.4	baseline optimal nonlinear analysis.
B&C		analysis-forecast cycle:
B	35 h at T=0.0	linear analysis at T=0
C	35 h at T=1.4	linear analysis at T=1.4, using a background forecast from expt B.
D	35 h at T=1.4	optimal nonlinear analysis.
E	35 h at T=1.4	linear analysis at T=1.4, using a background forecast from the T=0 background.
S	35 s at T=0.0 35 s at T=1.4	optimal nonlinear analysis of observations of wind-speed squared s = (U*U + V*V)/h*h.
VA	35 V at T=0.0 35 V at T=1.4	optimal nonlinear analysis of observations of a tracer V (with Coriolis parameter = 0).
VB&VC		analysis-forecast cycle:
VB	35 V at T=0.0	linear analysis at T=0
VC	35 V at T=1.4	linear analysis at T=1.4 using VB as background, of a tracer V (with Coriolis parameter = 0).
T	38 U at S=pi 38 V at S=pi	optimal nonlinear analysis.
TIN	38 U at S=pi 38 V at S=pi	optimal nonlinear analysis, with additional penalty on rapid variations.

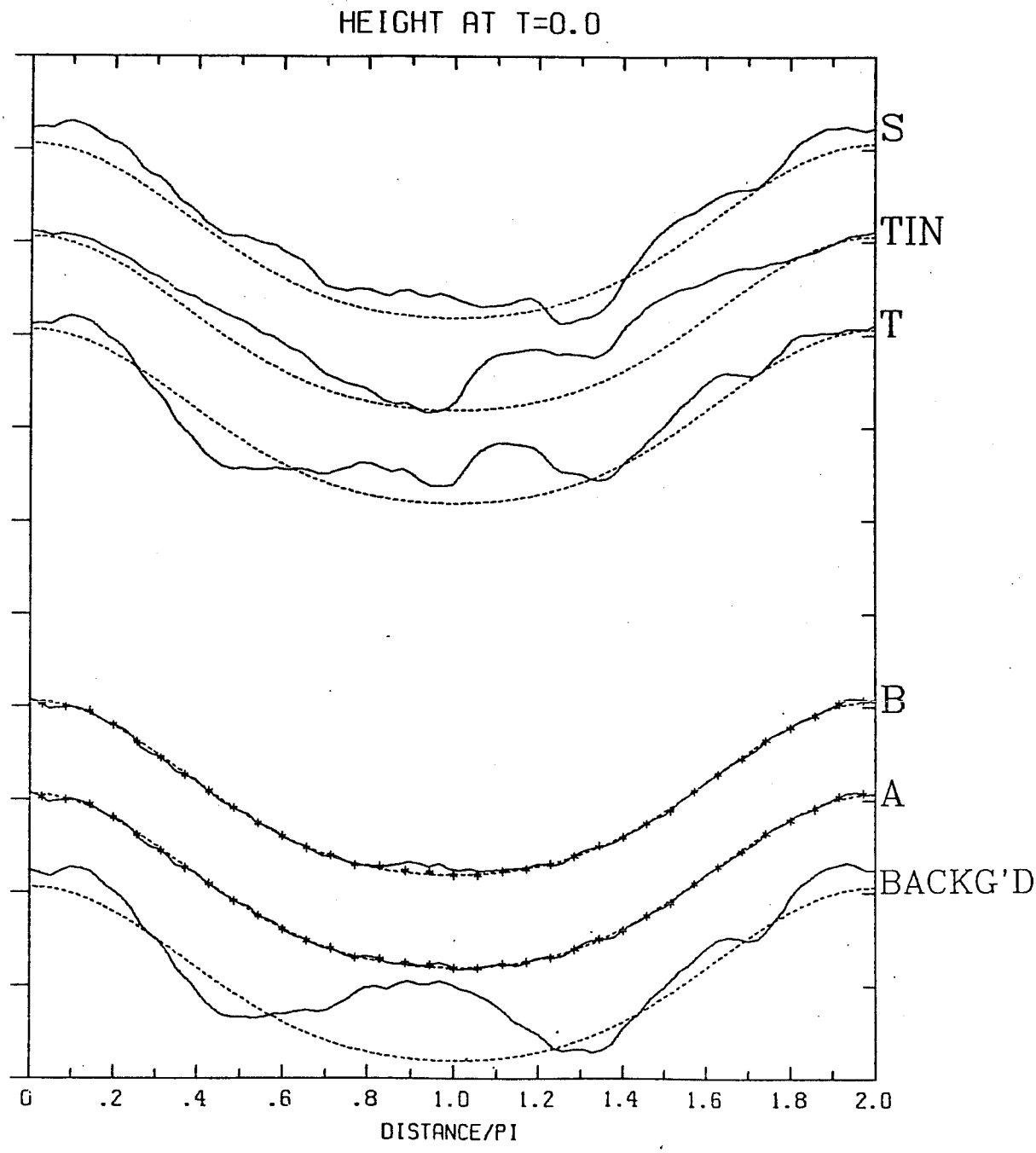

Fig.2. Height fields, plotted at time T=0.0. Curves for each experiment are displaced. Each curve has the corresponding curve from the "truth" shown dotted, and the observations used shown as *, with the height of the * showing 6 times the observational error and the width showing the horizontal gridlength. Graduation marks on the vertical scale are separated by one unit of (non-dimensional) h. Experiments shown are listed in table 1.

observations were generated as dotted lines, and observed values as asterisks whose height covers six times the observational error standard deviation, and whose width covers one model gridlength. The bottom curve in Fig.2 shows the background field, which in practice would come from a forecast from earlier analyses. The bottom curve in Fig.3 and Fig.4 shows a forecast from this background at T=1.4 and T=2.8, with the corresponding "true" field for comparison. To simulate the current operation situation, with most observations at main synoptic hours, we assume a uniform space-distribution of observations at T=0 and T=1.4. For simplicity of presentation we only use height obsevations for the first experiments. The curves labelled A in Fig.2 and Fig.3 show the optimal nonlinear analysis obtained using these observations. Fig.4 shows forecasts valid at T=2.8 which can be used to judge the extent that the information has been assimilated into the model. Experiments B and C together used the same observations in an analysis-forecast cycle. The curve labelled B in these Fig.2 shows the optimal analysis using only the observations available at T=0. With the covariances we have assumed, this is equivalent to a univariate OI of the height field at T=0, not altering the background winds. Forecasts from this are clearly inferior to experiment A. Experiment C used the forecast valid at T=1.4 from experiment B as background for an analysis, in space only, of the observations available at this time. The error covariances valid for the background at time T=0, were also used for the forecast errors. Experiment C is also clearly inferior to experiment A. At T=2.8 the hydraulic jump is as badly positioned as in the background, while experiment A has both the position and shape more nearly correct.

For some modern observing systems, observations are not at all synoptic, but rather spread evenly in time. Satellite temperature soundings were the first important example of this; an important example in the future will be fixed "profilers", giving a detailed time-history of the atmospheric profile at a few horizontal locations. Experiment T demonstrates that our method can make optimal use of such observations. To simulate a single profiler we generate u- and v-momentum observations

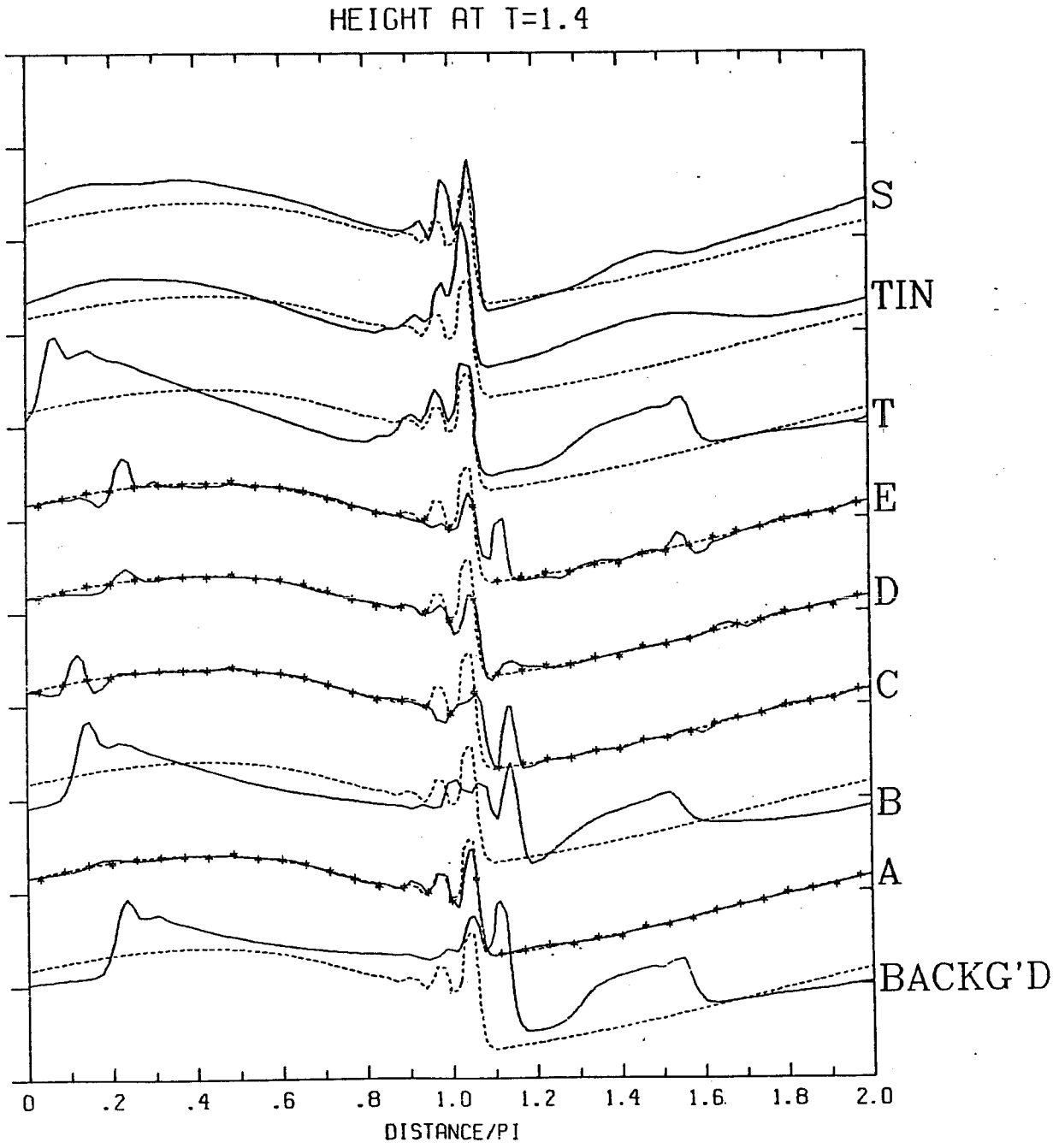

Fig.3. As Fig.2 for T=1.4.

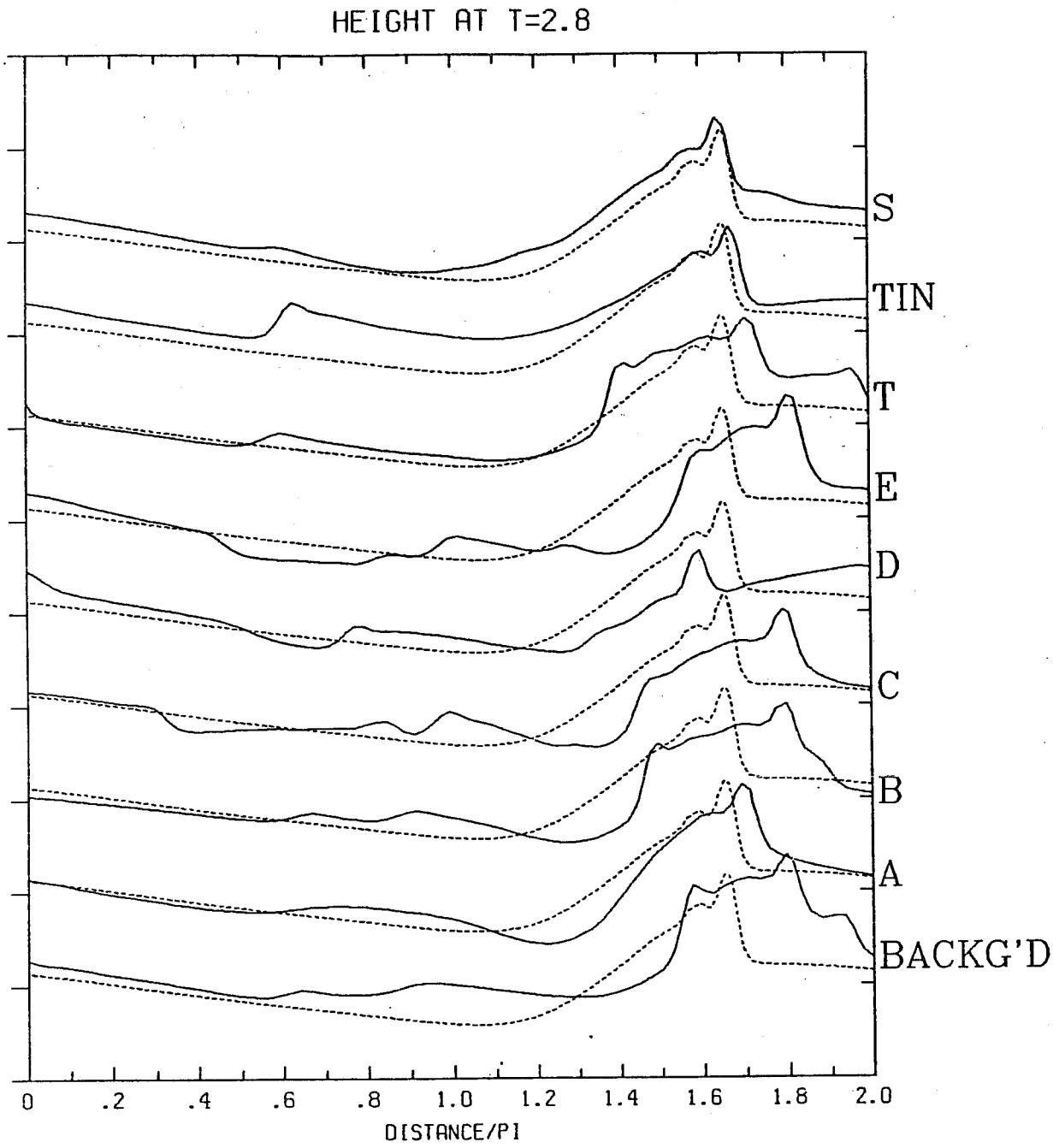

Fig.4. As Fig.2 for T=2.8.

equally distributed in time at a point in the middle of our horizontal domain. Fig.5 shows the u-momentum observations, and the time-evolution of the analysis at that point made using them. We can see clearly, near T=1.4, the passage through the observation position of the hydraulic jump. Since it is a propagating system it should be possible to deduce the horizontal structure of the jump near the profiler from this information. Examination of curve T in Fig.2 shows that this has indeed been done.

4b. Use of "indirect" observations.

Early experiments in four-dimensional data assimilation used rather ad hoc modifications of the "direct insertion" method, whereby observations were inserted into the forecast model state at and near their valid position and time. Using this method it is not clear how to use observations not directly related to model variables; observations are transformed to model variables before being inserted. If the objective analysis is itself couched as an inverse problem, with the transformation $\mathcal{K}$ from model space to observation space explicitly considered, then prior inversion is no longer necessary. As long as estimates of the observed parameters can be calculated from the model using a well-behaved differentiable function $\mathcal{K}$, the observations can be used in the analysis.

As an example of such a system, we simulate in experiment S observations of wind speed without direction information. We assume that observations are of parameter s, which can be calculated from the basic model parameters height h, and momentum components U and V by:

$$s = (U^2 + V^2) / h^2$$

This, together with the space- and time-interpolation from the nearest model points, constitutes our generalized interpolation operator $\mathcal{K}$. It is clearly nonlinear, particularly near the hydraulic jump where h can become very small. However its differential K is well defined, since h is constrained to be always greater than zero, and an optimal analysis should exist.

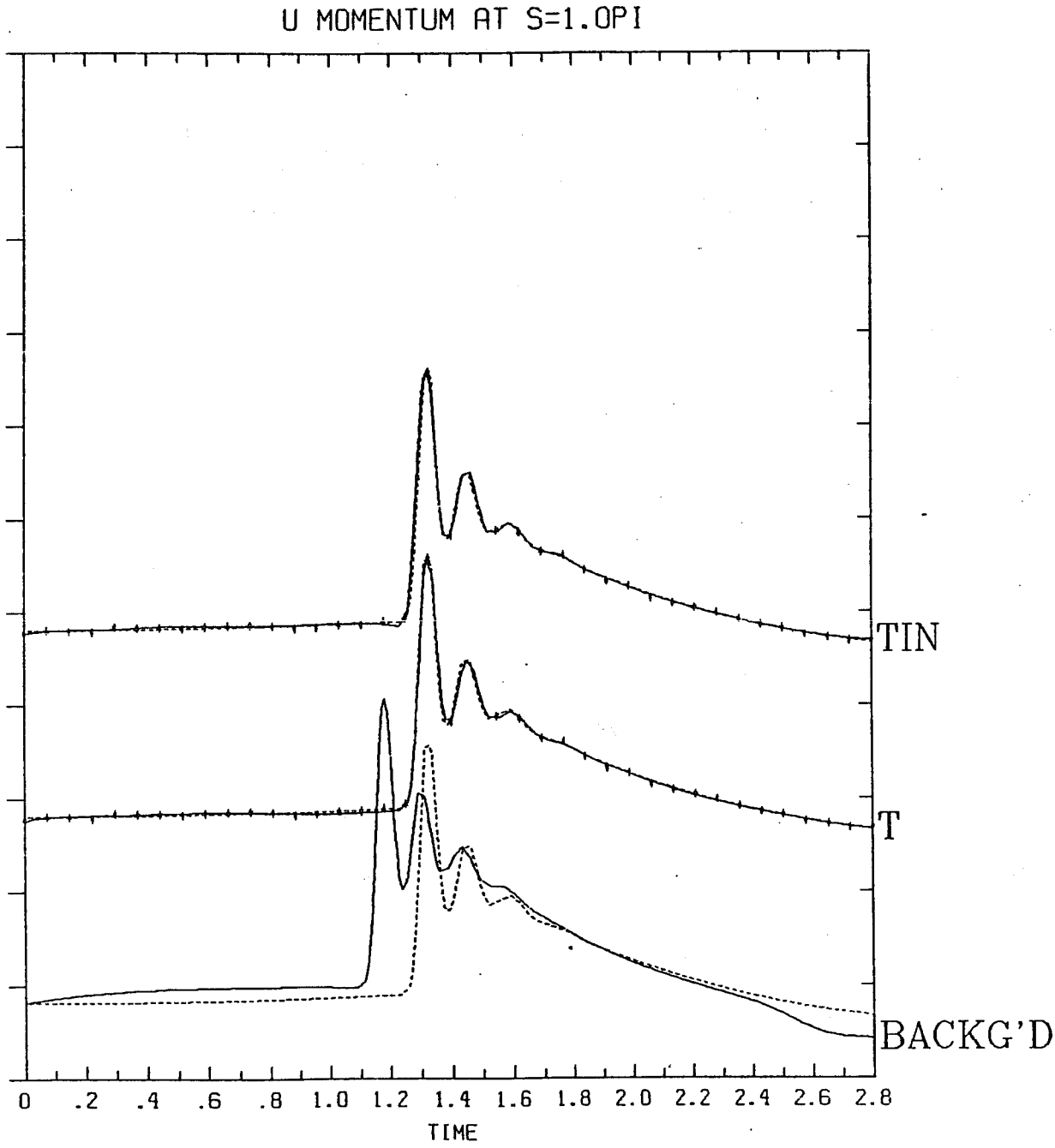

Fig.5. As Fig.2 for U at the centre of the grid, plotted against time.

Observations of s are distributed for experiment S at the same positions as the h observations of experiment A. Fig.6 and Fig.7 show that the analysis from experiment S fits these data closely, more so than that from experiment A. Although the experiment S analysis does not reproduce the detailed structure of the jump at T=1.4 as well as experiment A (Fig.3), it does give a better prediction of the jump's position and structure at T=2.8 (Fig.4).

Another type of data, used in subjective analyses and forecasts, but not useful in conventional updating data assimilation methods, is tracer information. A time-series of observations of a parameter advected by the wind field provides information about the wind field as well as about the advected field. We can simulate this in our simple model by setting the Coriolis parameter to zero, uncoupling V from h and U, and using V as a tracer. Fig.8 and Fig.9 show experiments with this system. In experiment VA, the full adjoint technique is used to modify h U and V to obtain the best fit to the V observations. In experiments VB and VC, h and U are kept at their background values, and an analysis-forecast cycle (as in experiments B and C) is used to update the V field. Useful corrections to the advecting velocity are made in experiment VA, so that its forecast V field at T=2.8 is better than that of experiment VC (Fig.10).

4c. Analysis of discontinuities and fronts.

For many years human analysts have had conceptual models of fronts, and have fitted these to rather sparse data, producing analyses with detailed structure such as sharp windshears, even when these were not well resolved by the observations. This process is nonlinear; the use the analyst makes of an observation depends on what he believes the meteorological situation to be, based in part on the observation value. Attempts to formalize such conceptual models and automate this analysis process have not been very successful. However we now have high-resolution NWP models which generate naturally very realistic looking frontal structures. These forecast models might be able to replace the conceptual models in a nonlinear analysis scheme, to produce analyses which are consistent both

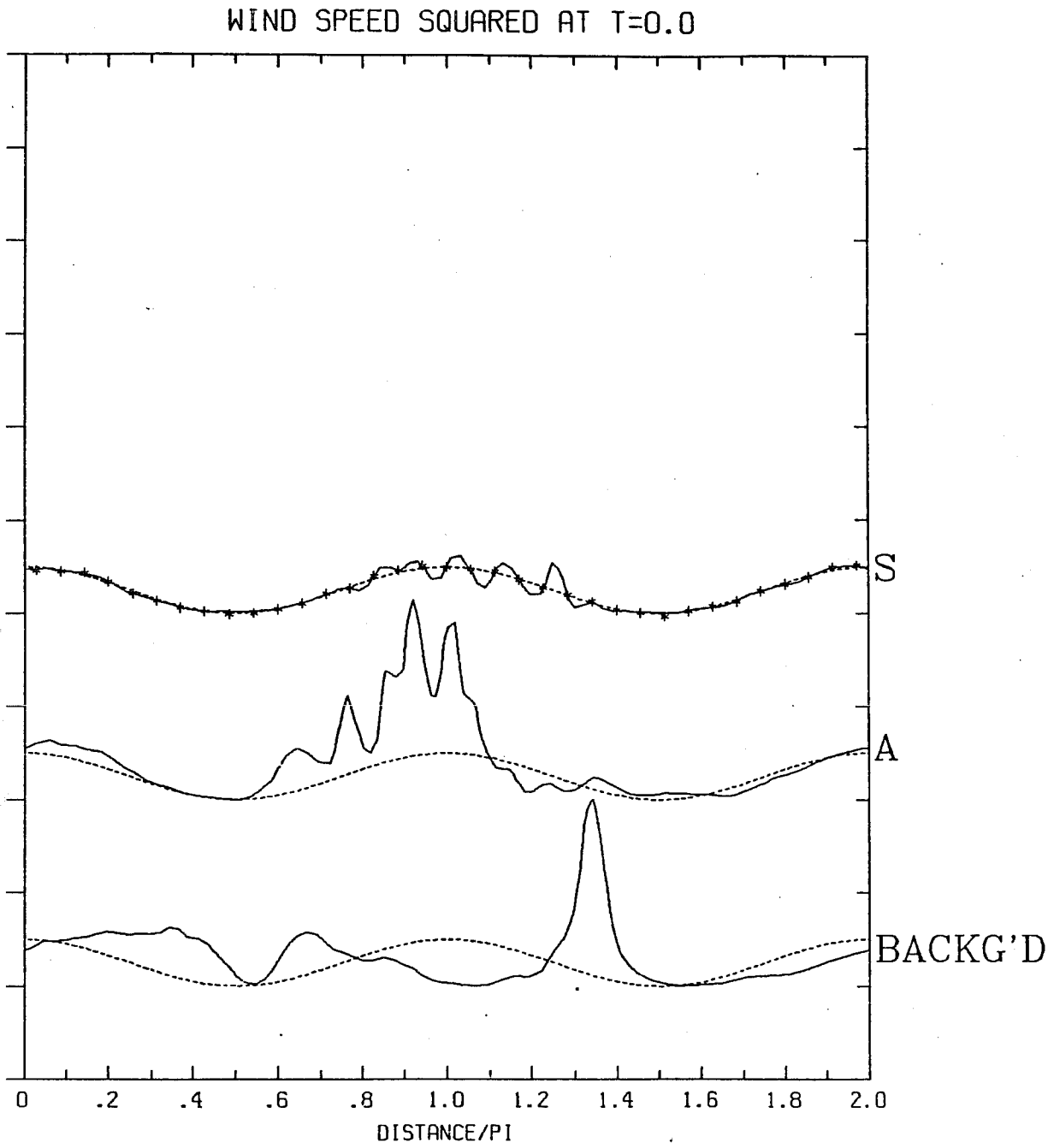

Fig.6. As Fig.2 for wind speed squared at T=0.0.

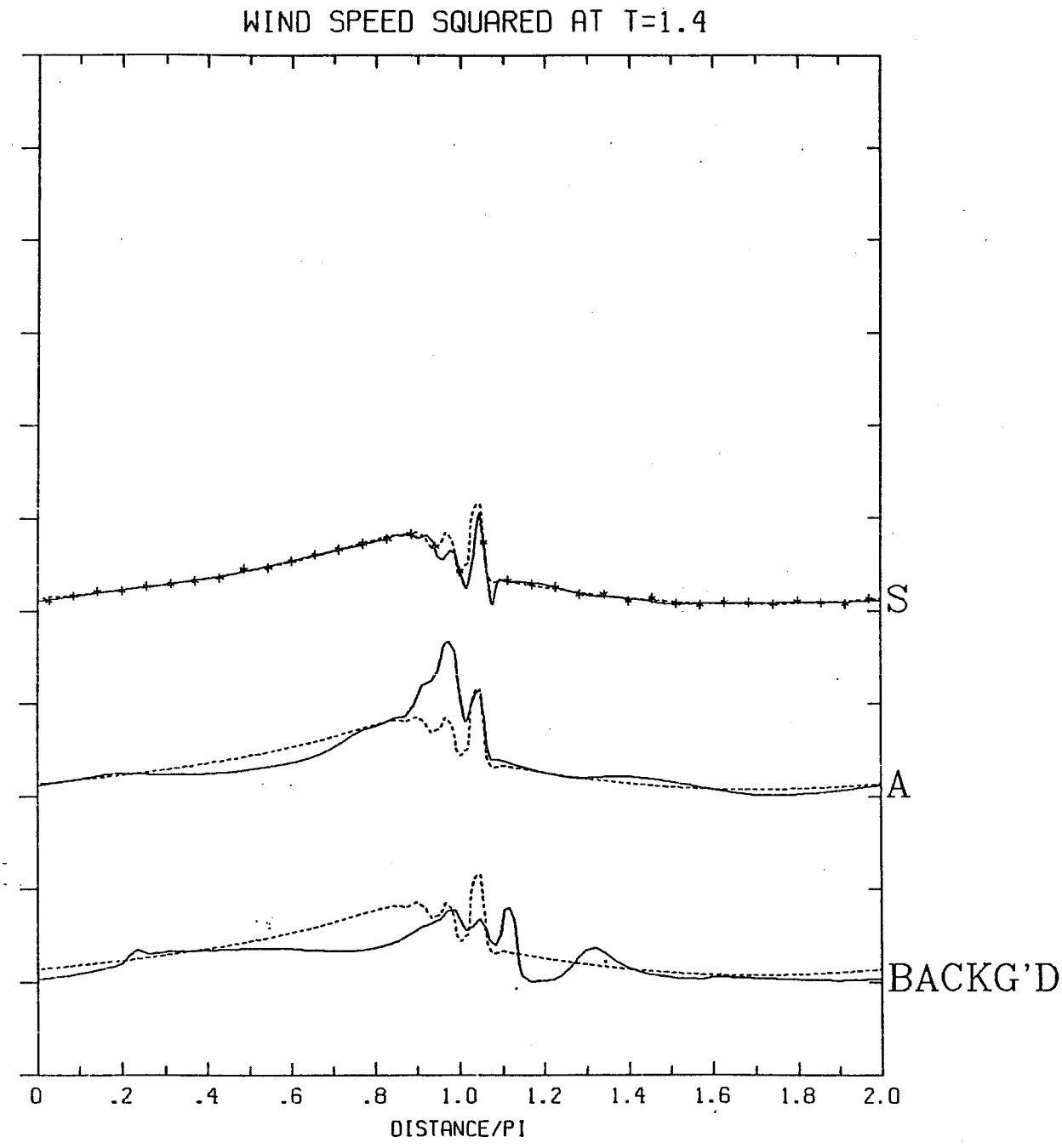

Fig.7. As Fig.2 for wind speed squared at T=1.4.

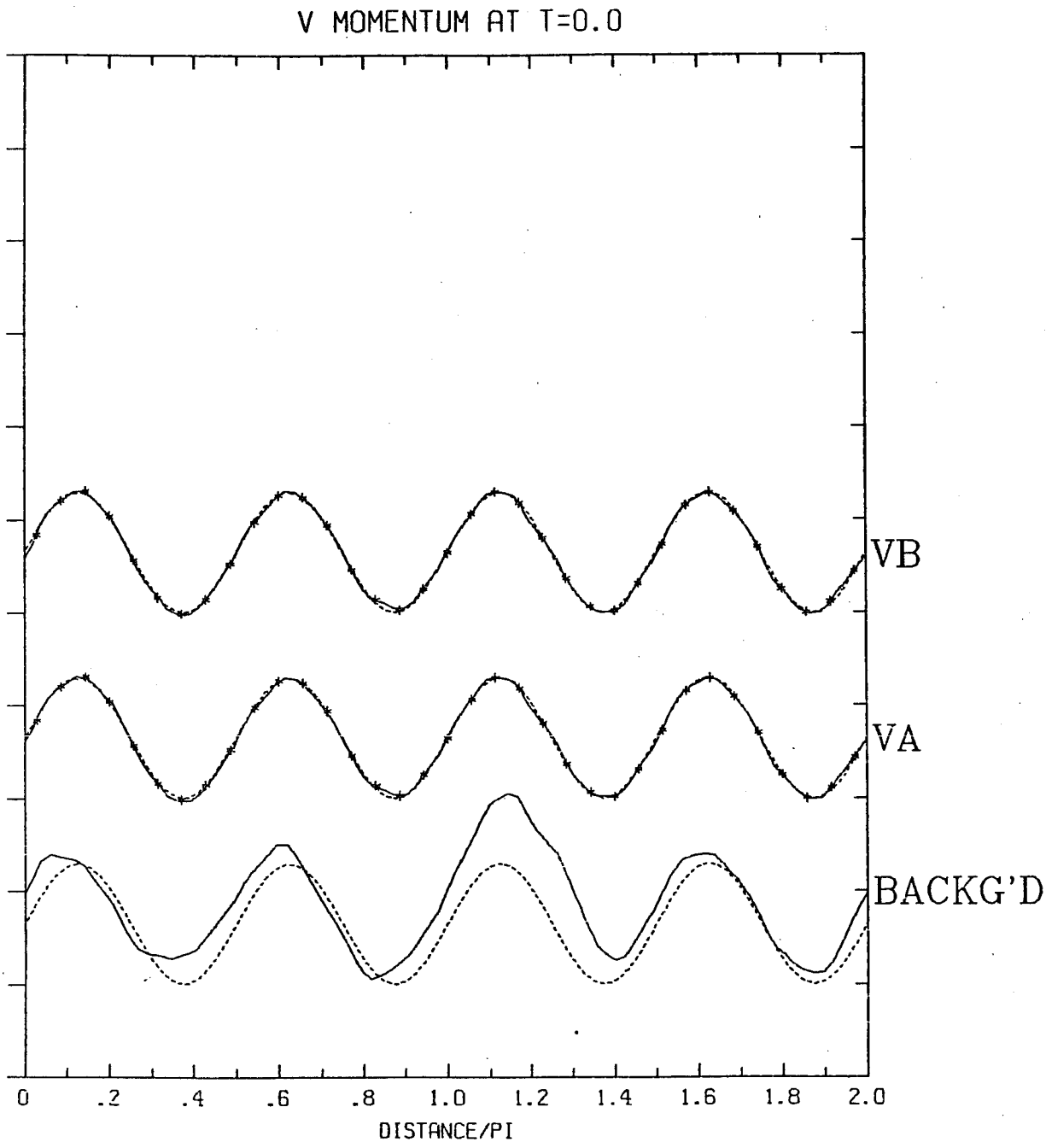

Fig.8. As Fig.2 for V at T=0.0, from experiments with zero Coriolis parameter, so that V was a simple tracer.

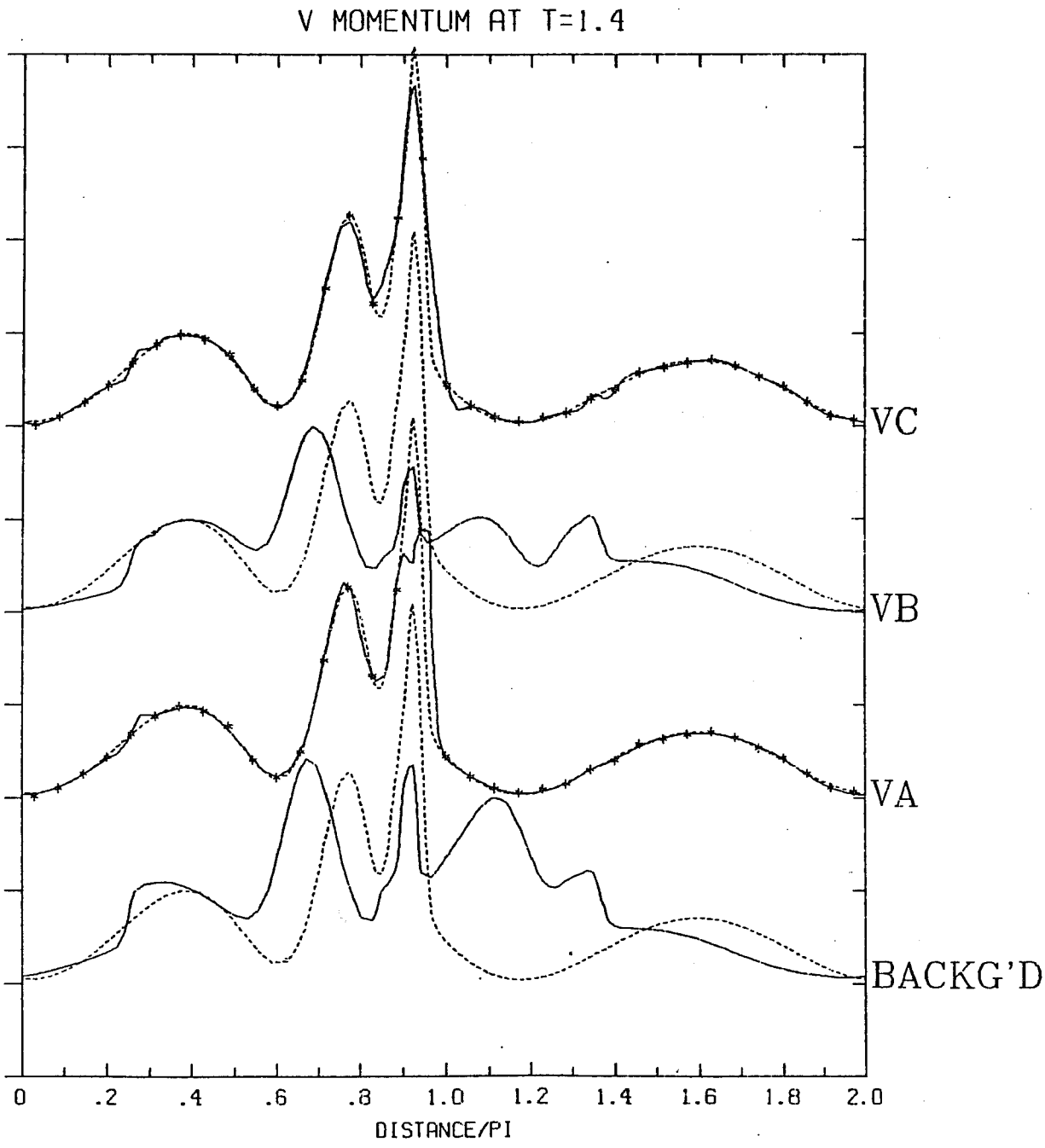

Fig.9. As Fig.2 for V at T=1.4, from experiments with zero Coriolis parameter, so that V was a simple tracer.

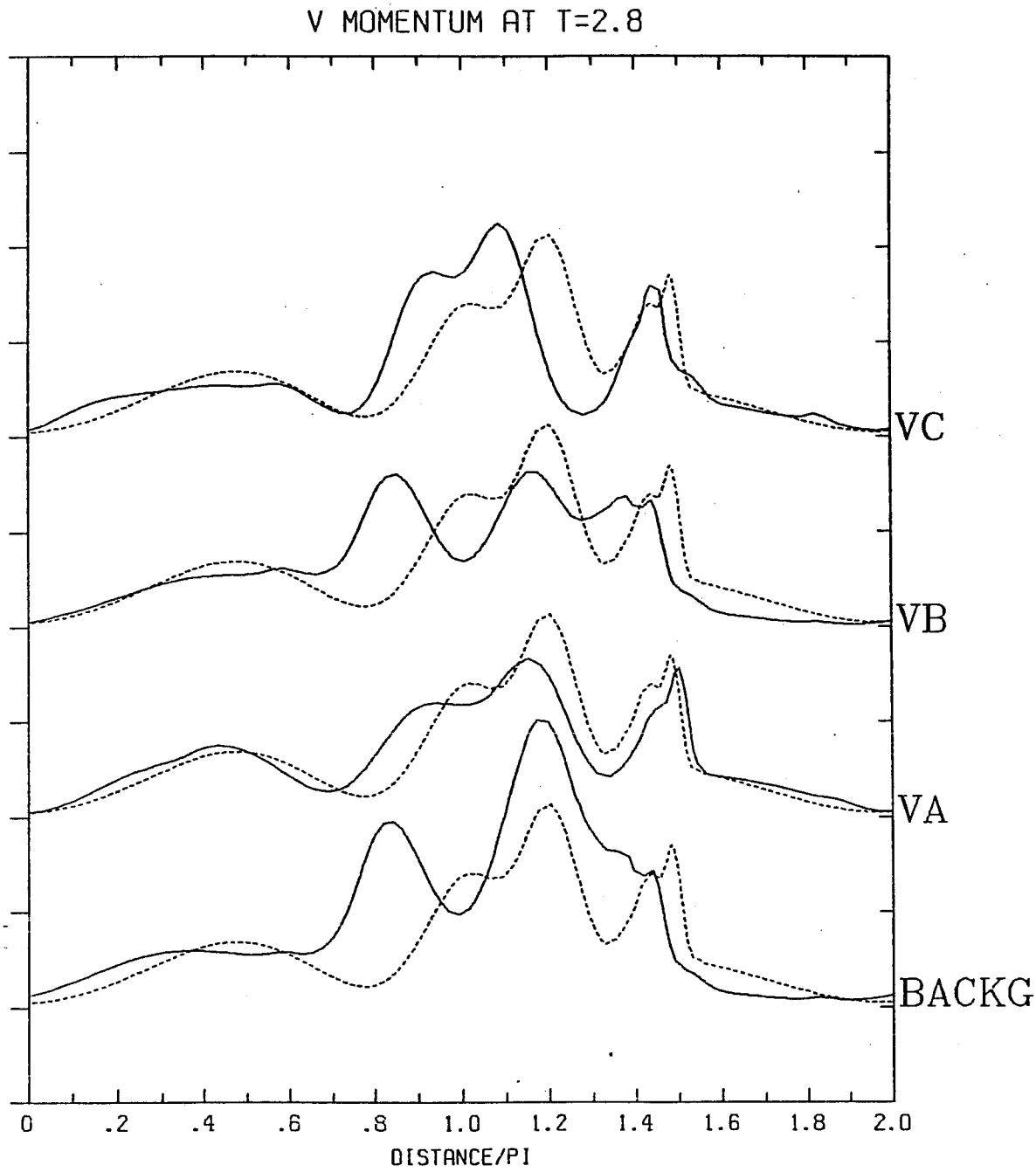

Fig.10. As Fig.2 for V at T=2.8, from experiments with zero Coriolis parameter, so that V was a simple tracer.

with the observations and with the model's dynamics. Given a background with a front somewhat misplaced, a human analyst will move it to fit available observations, while keeping basically the same structure. A linear analysis scheme like (1) cannot do this, as illustrated in Fig.3 curve E, which shows a linear, space-only, analysis of the observations at T=1.4 using the background shown in the bottom curve. The linear analysis has put a jump near the correct position, however its structure is incorrect, and "shadows" remain of jumps in incorrect positions in the background field. This can be compared with the nonlinear analysis shown in curve D, which used the same observations, but did a space- and time-analysis using as background the bottom curve in Fig.2. The nonlinear method has moved the hydraulic jump at T=1.4 to fit the data, while keeping its structure. Indeed this analysis is better than that from the analysis-forecast cycle (C), which used, linearly, twice as many observations. The beneficial nonlinear use of observations in experiment D comes from the use of the nonlinear forecast model to link T=0 with T=1.4. We still assume that errors at T=0 are large-scale and there is no coupling of scales. The method only gives an improvement in the analysis at T=1.4 insofar as the small scales at that time are determined by the larger scales at T=0. This is true for our chosen example; it is also probably true to a large extent for atmospheric fronts, whose position and structure depend on the larger scale forcing.

4d. Balance constraints, initialization.

We could incorporate a geostrophic balance constraint on the deviations of the analysis from the background. Such a constraint is linear, and only constrains the total analysis field if the background is itself balanced. It is sometimes desirable to incorporate a nonlinear constraint on the total analysis into the analysis process. Such a constraint can be justified if we have prior knowledge that the atmosphere is usually slowly varying. NWP models, if integrated for a sufficiently long time, also have this property. However since we are only using the model as a strong constraint over a short time-period, for which the model will also allow rapidly

varying gravity-wave modes, we must use other means to enforce the constraint. The simplest way is to add a penalty term which penalizes rapid variations. In order to have the same effect as nonlinear initialization, we implement this penalty on the time derivatives calculated by the NWP model during its first timestep at T=0. Experiment TIN incorporated such a penalty, with other details identical to experiment T. In current operational schemes a nonlinear initialization is often applied after the analysis is complete. This can significantly decrease the closeness of fit to the observations. We see in Fig.5 that experiment TIN has fitted the observations as closely as experiment T, while causing a slight reduction (not easily visible on the scale of Fig.5) in the initial rapid variation in U. Errors relative to the "true" field still exist in the large scale h field (Fig.3), since this is defined neither by the observations nor by the constraint, but the spurious jumps introduced by the background, and not altered except in position in experiment T, have been reduced by the initialization constraint. This causes the forecast structure at T=2.8 to be better (Fig.4).

4e. Non-Gaussian observational errors and quality control.

Unfortunately the observations available for routine NWP occasionally deviate by a large amount from the true value, because of gross error, either human, or in the instrument or communication system. Many more such errors occur than would be expected from the Gaussian distribution which describes the majority of errors. Traditionally such data are searched for and eliminated during a preliminary quality control step. Purser (1984) suggested that an alternative approach would be to consider the non-Gaussian error distribution directly in the analysis. A simple model for the distribution of gross errors has been put forward and tested by Lorenc and Hammon (1988). They postulated that there was a small probability of a gross error event occurring, and if it did occur the observed value had no useful information, but was equally likely to be any value within a range of plausible values. Fig.11 shows the penalty function for a single observation of this type by ln(1+g)), with the equivalent quadratic penalty function for a

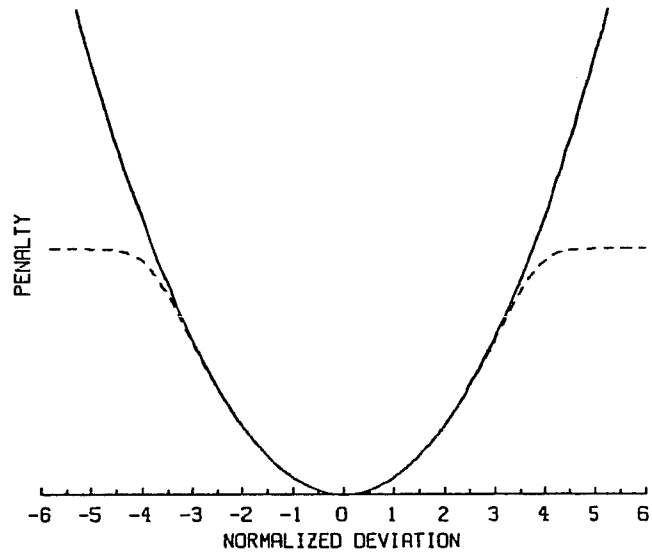

Fig.11. Solid line: quadratic (L2) penalty function for a single observation, plotted against the normalized deviation dy[i]/sqrt(O[i]). Dashed line: the equivalent penalty function derived assuming that the observation has a 5% chance of being useless because of a gross error.

pure Gaussian distribution for comparison. Near the observed value the functions are identical. However farther away the new function asymptotes to a plateau value, rather than continuing to increase. This has very important consequences for the total penalty function, since it makes the existence of multiple local minima much more likely. Our iterative minimization algorithms are designed to find any minimum, rather than the smallest, so they are therefore much less likely to find the most likely analysis.

Experiments illustrating this are shown in Fig.12 and Fig.13, and listed in Table.2. Particularly when using a non-Gaussian observational error probability distribution, the minimum located by the iterative solution method depends greatly on the first guess used to start the iteration. Thus experiment AQ fitted very few of the data, finding a local minimum near to its first guess, which was the background. This can be compared with experiment AQA, which found a lower minimum near the analysis of experiment A, fitting all the data. When one of the data values was arbitrarily increased by 0.5, to simulate a gross error, then starting from experiment A as first guess the datum was rejected while others were fitted (experiment AQGA in Fig.13). This is the result we might hope for from an ideal scheme. However the analysis (not shown) starting from the background field, fitted as few data as experiment AQ. Moreover the "bad" datum we generated was not completely implausible; experiment AG fitted it by moving the nearby peak in the background field. Starting with this analysis as first guess, experiment AQGAG continued to fit the "bad" datum, rejecting its neighbours instead. The minimum found in experiment AQGA was lowest of those shown, but a more complex search algorithm would be necessary to be sure it was the absolute minimum. It is clear that simply using a non-Gaussian penalty function does not relieve us from the need for complex logic and decision taking algorithms associated with traditional quality control methods.

Table 2. Experiments described in sections 4e.
Curves in Fig.12 & Fig.13 are labelled with Expt.
35 h observations were equally distributed in space at T=0.0 and at T=1.4, as in experiment A.
Experiments used a space grid of 128 points.

Expt.	P_{or}	erroneous data	First-guess
A	Gaussian	0	background
AQ	non-Gaussian	0	background
AQA	non-Gaussian	0	analysis from A
AG	Gaussian	1 at T=1.4	background
AQGA	non-Gaussian	1 at T=1.4	analysis from A
AQGAG	non-Gaussian	1 at T=1.4	analysis from AG

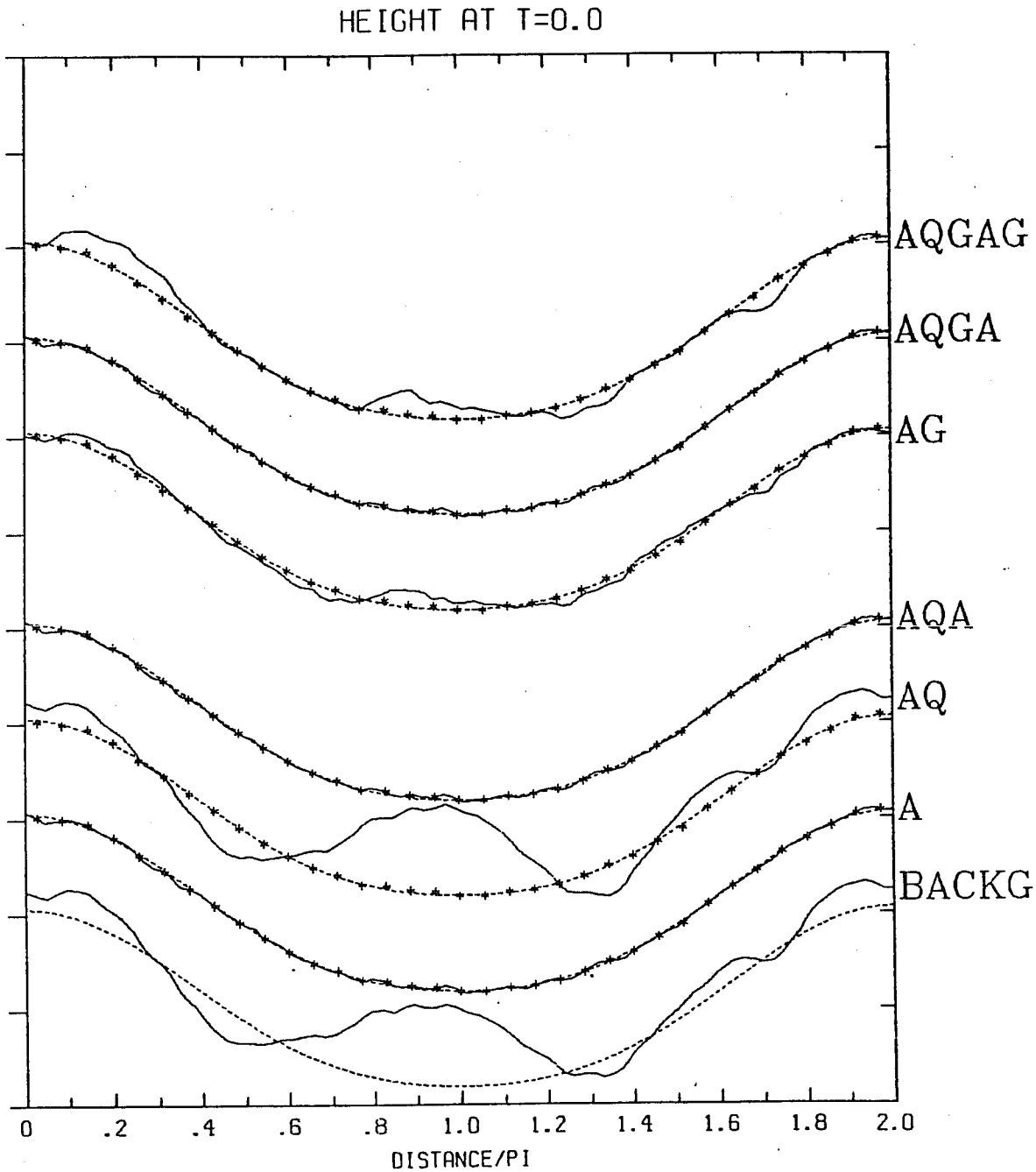

Fig.12. As Fig.2 for quality control experiments listed in Table.2. Non-Gaussian error distributions => 0 in label. A gross error => G in label. Analysis used as first guess indicated at end of label.

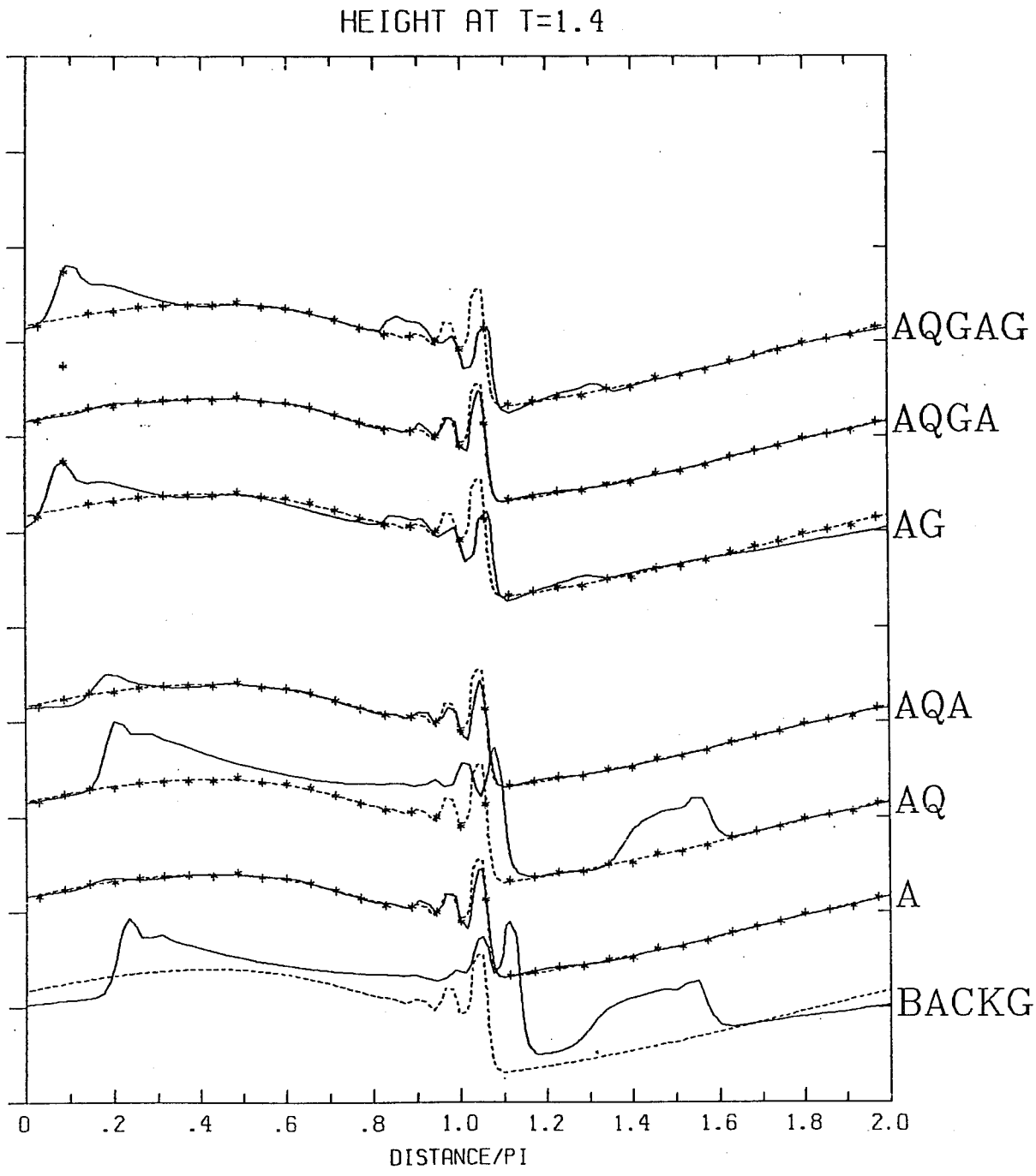

Fig.13. As Fig.3 for quality control experiments listed in Table.2. Non-Gaussian error distributions => Q in label. A gross error => G in label. Analysis used as first guess indicated at end of label.

6. DISCUSSION

We have presented examples of the solution of analysis equations which are nonlinear in the data values. That is, the weights given to observations depend on the analysed fields, not just on their positions and accuracy.

We demonstrated that:-

a. The nonlinear analysis method is able to use time-tendency information from observations better than the analysis-forecast cycle method which is normally used for data assimilation. It could also convert a time-sequence of data from a single observation location into useful information about the spatial structure of the field.

b. The data used in the analysis do not have to be transformed to the analysed parameters; it is only necessary to have a known method for calculation the observed parameters from those analysed. Thus observations of wind speed, without direction information, can be used, as can information about the advection of a tracer.

c. If the forecast model generates realistic structures for features like fronts, then the nonlinear method is capable of "moving" such a feature in the first-guess to fit the available data, even if the data do not resolve all the details of the feature. The resulting analysis is thus more detailed than a scale analysis of the observational distribution alone would lead one to expect.

d. An additional constraint, that the nonlinear evolution of the analysed field should be slow, can be incorporated as part of the analysis process. This improves the evolution of the subsequent forecast from the analysis, without greatly reducing the fit of the analysis to the observations.

e. Observations which are more likely to have large errors than would be expected from a normal distribution of instrumental errors, can be allowed for by specifying an appropriate non-Gaussian error distribution. If we make the

reasonable assumption that observations with such gross errors contain no useful information, then a limit is placed on the penalty function being minimized. This generates "plateau" regions, and greatly increases the difficulties in finding the best analysis if we do not have an accurate first-guess. Thus the complex logic required for a comprehensive quality control is not avoided. However, given a reasonable first-guess, the method does effectively ignore erroneous data.

REFERENCES

Lorenc,A.C. 1986 "Analysis methods for numerical weather prediction." Quart. J. Roy. Met. Soc., **112**, 1177-1194

Lorenc,A.C. 1988a "Optimal nonlinear objective analysis" Quart. J. R. Met. Soc., **114**, 205-240

Lorenc,A.C. 1988b "Iterative approximations to objective analysis." ECMWF Seminars on Data assimilation and the use of satellite data. September 1988. (this volume).

Lorenc,A.C., and Hammon,O. 1988 "Objective quality control of observations using Bayesian methods - Theory, and a practical implementation." Quart. J. Roy. Met. Soc., **114**, 515-543

Parrett,C.A., and Cullen,M.J.P. 1984 "Simulation of hydraulic jumps in the presence of rotation and mountains." Quart. J. R. Met. Soc., **110**, 147-165

Purser,R.J. 1984 "A new approach to the optimal assimilation of meteorological data by iterative Bayesian analysis." Preprints, 10th conference on weather forecasting and analysis. Am. Met. Soc., 102-105