U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANOGRAPHIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
NATIONAL METEOROLOGICAL CENTER

OFFICE NOTE 397

OPTIMAL AVERAGING OF METEOROLOGICAL FIELDS

Lev S. Gandin
UCAR

JULY 1993

1. INTRODUCTION

Although every measurement, including meteorological ones, results in a value averaged to some extent both in space and in time, a majority of measurements, or observations, particularly traditional observations at meteorological stations, are usually refered to, and considered as, giving instantaneous point values. At the same time, it is practically always assumed, explicitly or implicitly, that every result of such measurement is applicable, or *representative*, not only for the point and time of the measurement, but also for some spatial, particularly horizontal, domain surrounding the observation point, as well as for some interval of time.

The degree of this representativeness is, of course, different for different parameters depending, first of all, on their spatial and temporal variability: the higher the variability, the smaller is the representativeness domain, and the more often, in space and time, should be therefore the observations performed in order to assure a given degree of accuracy everywhere between the observations. Expressed quantitatively, in terms of the root mean square (RMS) interpolation error, this requirement gives a rationale for the planning of observation station networks knowing the statistical structure of the meteorological parameter in question (e.g., Gandin, 1970).

There exist, however, meteorological parameters with such high variability that the above requirement cannot be satisfied with any reasonable network density. The daily precipitation amount is the most known among such parameters. There are many others, like evaporation, snow cover, radiative heat fluxes, etc. For these parameters, the requirement of sufficient accuracy at any point within an area should be, and usually is, replaced by a weaker requirement that the values averaged over the area can be determined with sufficient accuracy.

The area-averaged values of parameters with high spatial variability, like precipitation and evaporation (and runoff) averaged over a river basin, are of primary

practical value. That is why several methods to estimate an area-mean value from several point measurement data were developed in hydro-meteorology comparatively long ago. A systematic review of these methods may be found in the monograph "Averaging of meteorological fields" by R. Kagan (1979, hereafter refered to as K79).

Development of remote observations, particularly of those from meteorological satellites, has resulted in increased importance of the spatial averaging problem. Although the spatial resolution of modern remote sounding data is quite high, these data certainly reflect values of meteorological parameters averaged over some domains. A kind of averaging of point data is therefore needed if we wish to examine the agreement between them and remote sounding information.

The main applications of temporally and spatially averaged data are, however, caused by the increase in accuracy due to the averaging, which is particularly important for long-range weather forecasts (often called climate forecasts) and even more for investigations of the global climate change.

The problem of climate change, also known as that of global warming, attracts very much attention nowadays. There are hundreds of publications devoted to this problem. Still, many of these publications consider some specific, more or less localized, phenomena presenting them as demonstrating already existing global warming (such events are often refered to as "signatures" or even "fingerprints" of the global warming). Needless to say, these "fingerprints", however interesting they may be by themselves, cannot be considered as a proof of the presence (or absence) of the global warming. The only way to empirically investigate the problem should be connected with attempts to distinguish a small but persistent climate change on the background of much more intensive everyday weather variability. In order to achieve this aim, it is necessary to apply the averaging, both in time and in space, to a largest reasonable extent.

Some investigations of that kind have been performed with some surface observation data, mainly with those on the surface air temperature. Several authors computed and analysed time series of monthly mean temperature averaged over various latitudinal belts, or even over the Northern hemisphere and found indications that a kind of global warming is already taking place (e.g., Budyko, 1982). Many far-reaching conclusions, predictions and proposals have been made on the basis of these findings.

The situation is, however, not so obvious as it might seem. Before making any conclusion, it is necessary to evaluate not only the change itself, but also the root-mean-square (RMS) error of this evaluation. This is not a simple task. Specifically, the well-known "square root theorem" (stating that the RMS error in an average of n values is inversly proportional to the square root of n) is not applicable here: it holds only for averaging of *independent* data, while the values of a meteorological parameter like the surface air temperature at various points or at various times may be strongly interdependent. If, nevertheless, the theorem is applied to the problem under consideration, it gives rather unrealistic, over-optimistic estimates of the averaging accuracy.

A new approach to this problem has been developed by R. Kagan and his colleagues about two decades ago. It is presented in a systematic way in K79. The approach uses information on so-called statistical structure of the meteorological parameter in question, that is, on its variance, correlation function and RMS observation error. As long as the averaging over an area is linear with respect to the point values, as it practically always is, Kagan's method allows us to estimate the RMS error of averaged values as a function of the averaging method, of geometry (size and configuration of the area, number and pattern of the points), and of the statistical structure.

Moreover, it is possible to find, for any given geometry and statistical structure, such a linear combination of point values which approximate the area mean with the

minimal RMS error. This procedure, called the optimal averaging, is in many respects analogous to the so-called optimal (or optimum, or statistical) interpolation widely applied in the objective analysis of meteorological fields. Due mainly to this application, the optimal interpolation method is well known by the meteorological community, it is described in recent monographs by J. Thiebaux and M. Pedder (1987) and by R. Daley (1991). At the same time, the optimal averaging method remains almost uknown among American specialists, as does the Kagan's approach in general. That is due to the fact that almost all publications on the optimal averaging, including K79, are available only in Russian.

New possibilities to perform the area averaging of various meteorological parameters on a regular basis are emerging in connection with the Climate Data Assimilation System (CDAS) designed now at NMC. Collecting most complete sets of data and being not restricted in time by operational requirements, CDAS also makes it possible to average deviations from the forecast first guess instead of those from climatology, as it has been done before, and thus to achieve a higher accuracy. Even more important from this point of view is the Reanalysis project developed in parallel with CDAS (Kalnay and Jenne, 1991). Its aim is to perform anew objective analyses of past data using modern methodology of the data assimilation like that of CDAS.

The design of optimal averaging is considered an integral part of CDAS and Reanalysis projects, as is the work reported in this Office Note. Its purpose is mainly methodological: to analyse main properties, first of all the accuracy, of the optimal averaging (OAv) under various circumstances and to compare the OAv with other procedures, like the optimal interpolation and the arithmetic averaging. In order to be able to consider numerous effects in a most distinct way, a rather simplified OAv model - we call it the "toy example" - is used throughout this investigation. It allows us to perform

various integrations in analytical or, to be more exact, in a "quasi-analytical" form instead of using numerical methods.

The text of this Office Note is organized as follows: General equations expressing the RMS error of approximation of an area-mean value by any linear combination of point values are presented in Section 2, and OAv is introduced as a special case. Section 3 describes the "toy example". Equations for the variance of the area-mean and for the cross-correlation between it and a point value are derived and used in this section. These equations are applied to the simplest case of using one point to characterize the area-mean, i.e., to the problem of representativeness, in Section 4.

The most important, multi-point, situation is considered in some details in Section 5. Influence of various factors, like the area size, the number and disposition of points etc., is analysed, and conclusions are drawn concerning rational ways of specifying these factors. Section 6 is devoted to the influence of practically inevitable violations of optimality - those caused by the incomplete knowledge or approximation of the underlying statistics. The final Section 7 contains general conclusions as well as recomendations concerning the practical application of OAv in CDAS and Reanalysis.

2. AVERAGING ACCURACY AND STATISTICAL STRUCTURE.

As mentioned above, there exist many methods to estimate area mean values of meteorological and hydrological parameters from observed point data on these parameters. However different from each other these methods can be, each of them finally results in presenting the area mean as a linear combination of observed values. More exactly, this is true for deviations F of parameter in question from some background field, usually for deviations from climatological norms (the so-called anomalies). Various area averaging methods differ from each other only in how they specify the so-called averaging weights, i.e., the coefficients of the linear combination. This means that the value of F, averaged over some domain (S),

$$A = \frac{1}{S}\iint_{(S)} F(x,y)dxdy, \tag{1}$$

where x and y are horizontal coordinates and S is the area of the domain (S), is approximated by

$$\hat{A} = \sum_{i=1}^{n} w_i \tilde{F}_i, \tag{2}$$

where n is the number of points used in the approximation,

$$\tilde{F}_i = F_i + \delta_i \tag{3}$$

is an observed value of F_i and δ_i is a random observation error at point i. It will be assumed throughout most of this paper that random errors are symmetrically distributed

$$\overline{\delta}_i = 0, \tag{4}$$

where the overbar denotes statistical averaging, that they are non-correlated at different points

$$\overline{\delta_i \delta_j} = 0 \tag{5}$$

and that their root-mean-square value (the RMS observation error) Δ is constant within the domain

$$\overline{\delta_i^2} = \Delta^2 \qquad (i = 1,2,\ldots,n). \tag{6}$$

As to the random function F itself, we assume that it is statistically homogeneous and isotropic (with respect to second-order moments) within the domain, i. e, that its variance $D^2 = \overline{F^2}$ is constant

$$\overline{F_i^2} = D^2 \tag{7}$$

and that its autocorrelation function μ depends only on distance

$$\frac{\overline{F_i F_j}}{D^2} = \mu(r_{ij}), \tag{8}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is the distance between points i and j.

The area-averaging accuracy, that is, the accuracy of approximating the area-mean value A (1) by the linear combination $\hat{A}$ (2), may be characterized by the RMS averaging error E, or by the averaging error variance

$$E^2 = \overline{(\hat{A} - A)^2}. \tag{9}$$

Substituting expressions (1)-(3) into (9) and using assumptions (4)-(8), it is easy to obtain the following equation

$$E^2 = D^2 \sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \mu(r_{ik}) + \Delta^2 \sum_{i=1}^{n} w_i^2 + \frac{1}{S^2}\overline{\left(\iint_{(S)} F(x,y)\,dxdy\right)^2} - \sum_{i=1}^{n} 2w_i \frac{1}{S}\iint_{(S)} \overline{F(x,y)F(x_i,y_i)}\,dxdy. \tag{10}$$

The term

$$\frac{1}{S^2}\overline{\left(\iint_{(S)} F(x,y)dxdy\right)^2} = \frac{1}{S^2}\iint_{(S)(S)}\iint \overline{F(x,y)F(x',y')}dxdydx'\,dy' \tag{11}$$

in (10) represents the variance of the area-averaged value, $B^2 = \overline{A^2}$. Using (8), one can express (11) in the form

$$B^2 = D^2\frac{1}{S^2}\iint_{(S)(S)}\iint \mu\left(\sqrt{(x-x')^2+(y-y')^2}\right)dxdydx'\,dy'. \tag{12}$$

Analogously, the expression

$$\frac{1}{S}\iint_{(S)} \overline{F(x,y)F(x_i,y_i)}dxdy \tag{13}$$

in the last term of (10) represents the covariance between the point value F_i and area-mean value, $C = \overline{AF_i}$, and may be rewritten as

$$C = D^2\frac{1}{S^2}\iint_{(S)} \mu\left(\sqrt{(x-x_i)^2+(y-y_i)^2}\right)dxdy \tag{14}$$

Using (12) and (14), one can rewrite equation (10), after dividing it by D^2, in the form

$$\begin{aligned}\varepsilon^2 = &\sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \mu(r_{ik}) + \eta^2\sum_{i=1}^{n} w_i{}^2 \\ &-\frac{2}{S}\sum_{i=1}^{n} w_i \iint_{(S)} \mu\left(\sqrt{(x-x_i)^2+(y-y_i)^2}\right)dxdy \\ &+\frac{1}{S^2}\iint_{(S)(S)}\iint \mu\left(\sqrt{(x-x')^2+(y-y')^2}\right)dxdydx'\,dy'\end{aligned} \tag{15}$$

or in a short form, using the notations above,

$$\varepsilon^2 = \sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \mu(r_{ik}) + \eta^2\sum_{i=1}^{n} w_i{}^2 - 2\sum_{i=1}^{n} w_i\zeta_i + \beta^2 \tag{16}$$

Here

$$\varepsilon^2 = E^2 / D^2,\ \ \eta^2 = \Delta^2 / D^2,\ \ \beta^2 = B^2 / D^2 \ \text{ and } \ \zeta_i = C_i / D^2 \tag{17}$$

are, respectively, variances of averaging error, of random observation error and of area-averaged value, and covariance between area-averaged and point values, normalized by the variance of parameter F itself. Note that the cross-correlation function between the area-averaged and point values

$$\gamma = C / (BD) = D / B\zeta \tag{18}$$

differs from the normalized cross-covariance ζ.

Equation (16) (or (15)) allows us to evaluate the RMS error of approximation of the area-averaged value by any linear combination of observed point values, i.e., by an expression (2) with any given weights w_i. One can instead determine the vector of weights in such a way that ε becomes a minimum as compared with its values with any other choice of the weight vector. Necessary conditions of minimum

$$\partial\varepsilon / \partial w_i = 0 \qquad (i = 1, 2, \ldots, n),$$

as applied to (16), result in a system of linear equations

$$\sum_{i=1}^{n} \mu(r_{ik}) w_k + \eta^2 w_i = \zeta_i \qquad (i = 1, 2, \ldots, n) \tag{19}$$

for weights w_i. The averaging with weights found by solving the system (19) is called the optimal averaging. Using (19), it is possible to obtain from (16) a simpler equation for the (normalized) variance of the optimal averaging error:

$$\varepsilon^2 = \beta^2 - \sum_{i=1}^{n} \zeta_i w_i \ . \tag{20}$$

It is worthwhile to mention that the quantity ε defined in (17) expresses the ratio of the RMS area-averaging error E to the standard deviation D of *point* values of parameter F. Along with this quantity, or even instead of it, one may use the ratio of the averaging error E to the standard deviation B of area-mean values

$$\lambda = E / B = \varepsilon / \beta. \qquad (21)$$

We will call λ the relative RMS averaging error, in contrast with the "absolute" RMS error ε. The variance B of area-averaged values is always less (or, to be more exact, never larger) than the point variance D, and λ is therefore always larger than ε. Which of the two averaging accuracy measures, ε or λ , should be of primary interest, depends on the problem under consideration. The degree of our confidence in area-averaged values themselves is better described by λ, because it relates the averaging accuracy to the variability of averaged values. If, however, we want to use the area averaging to obtain results less influenced by a small-scale variability than the point values are, then ε should be prefered, because it relates the area averaging accuracy to the variability of point values.

It may be also mentioned that equations (16),(19) and (20) are applicable to approximation of any quantity depending on a meteorological parameter F, not necessarily of its area-averaged value, by a linear combination of observed values of this parameter. What is different for various problems of this kind are the equations expressing the variance of the quantity in question and the covariance between it and F in terms of the autocorrelation function of F. Particularly, equations (19) and (20) are valid for the optimal interpolation of F into some point 0. In this case $\beta = 1$ and $\zeta_i = \mu(r_{i0})$. The optimal interpolation may be also considered as a limiting case of the optimal averaging when the area size tends to zero.

Equations (19) and (20) are also applicable to a more general situation of averaging, when we need to obtain an area-mean value not of parameter F itself but of some of its differential characteristics, like vorticity or divergence. Possible applications of this generalization of the optimal averaging will be discussed briefly in Section 7.

3. SIMPLIFIED AREA-AVERAGING MODEL (THE "TOY EXAMPLE")

As was mentioned in the Introduction, the purpose of the investigation reported in this Office Note was to provide illustration of main properties of the optimal averaging as functions of its various parameters, as well as to compare the optimal averaging with some other procedures, like the arithmetic averaging and the optimal interpolation. It is convenient to do so using a rather simplified model so that various effects will appear in their "pure" form, rather than complicated by the influence of other effects. It is also important that computations with this "toy example" do not require numerical multi-dimensional quadratures and may be therefore easily performed for large amounts of cases.

There are two main assumptions resulting in our toy example: a) The autocorrelation function $\zeta_i = \mu(r_{i0})$ is the so-called "Gaussian" (or "negative square exponential") one:

$$\mu(r) = \exp(-r^2 / r_0^2). \tag{22}$$

Here r is the distance between two points and r_0 is the correlation radius understood as the distance where the correlation is equal to 1/e. Without loss in generality, one can put $r_0 =$ 1 which simply means that r_0 is taken as the unit in measuring distances. We will call distances measured this way relative distances. For them,

$$\mu(r) = \exp(-r^2). \tag{23}$$

b) The averaging area (S) is a circle on the plane (which implies that distances under consideration are small as compared with the Earth's radius). Its (relative) radius is denoted R_0. Polar coordinates (r, φ) with the origin in the area's center will be used, so that the distance d_{12} between two points (r_1, φ_1) and (r_2, φ_2), is

$$d_{12} = r_1^2 + r_2^2 - 2r_1 r_2 \cos(\varphi_1 - \varphi_2). \tag{24}$$

It follows from the simplifying assumption b) that the (normalized) cross-covariance ζ between the area mean value

$$(1 / \pi R_0^2) \int_0^{2\pi} \int_0^{R_0} F(r, \varphi) r dr d\varphi \tag{25}$$

and the value at a point is isotropic, depending only on the distance R between the area center and the point in question or, in other words, on the radius-vector of the point. Namely,

$$\zeta(R,R_0)=\frac{1}{\pi R_0^2 D^2}\int_0^{2\pi}\int_0^{R_0}\overline{F(R,0)F(r,\varphi)}rdrd\varphi$$
$$=\frac{1}{\pi R_0^2}\int_0^{2\pi}\int_0^{R_0}\exp\left[-(R^2+r^2-2Rr\cos\varphi)\right]rdrd\varphi \tag{26}$$

Using the integral presentation for the modified Bessel function of first kind and zero order

$$I_0(z)=\frac{1}{\pi}\int_0^{\pi}\exp(-z\cos\varphi)d\varphi, \tag{27}$$

one can transform (26) to

$$\zeta(R,R_0)=\frac{2}{R_0^2}\exp(-R^2)\int_0^{R_0}\exp(-r^2)I_0(2Rr)rdr, \tag{28}$$

or, by means of the series for I_0

$$I_0(2z)=\sum_{k=0}^{\infty}\frac{1}{(k!)^2}z^{2k}, \tag{29}$$

$$\zeta(R,R_0)=\frac{1}{c}\exp(-b)\sum_{k=0}^{\infty}\frac{b^k}{(k!)^2}\int_0^c\exp(-z)z^k dk, \tag{30}$$

where

$$c=R_0^2\ ;\ b=R^2. \tag{31}$$

A general expression for the integral in (30) may be written as

$$\int_0^c\exp(-z)z^k dk=k!\left[1-\exp(-c)\sum_{i=0}^{k}\frac{c^i}{i!}\right], \tag{32}$$

and substitution of (32) into (30) gives, finally

$$\zeta(R,R_0)=\frac{1}{c}\left[1-\exp(-(c+b))\sum_{k=0}^{\infty}\frac{b^k}{k!}\sum_{i=0}^{k}\frac{c^i}{i!}\right]. \tag{33}$$

The series in (33) converges fast enough to allow very fast and efficient computations by a simple computer code.

Analogous transformations may be applied to the expression for the (normalized) variance of an area-averaged value

$$\begin{aligned}\beta^2(R_0)&=\overline{\left[\frac{1}{DS}\int_0^{R_0}\int_0^{2\pi}F(r,\varphi)rdrd\varphi\right]^2}\\&=\frac{1}{\pi^2R_0^4D^2}\int_0^{R_0}\int_0^{R_0}\int_0^{2\pi}\int_0^{2\pi}F(r_1,\varphi_1)F(r_2,\varphi_2)r_1r_2dr_1dr_2d\varphi_1d\varphi_2.\end{aligned} \tag{34}$$

Transformation of variables

$$(\varphi_1+\varphi_2)/2=\bar{\varphi}\ ;\ \varphi_1-\varphi_2=\varphi \tag{35}$$

(with Jacobian equal to 1) gives

$$\begin{aligned}\beta^2(R_0)&=\frac{2}{\pi R_0^4}\int_0^{R_0}\int_0^{R_0}\int_0^{2\pi}\mu\left[\sqrt{r_1^2+r_2^2-2r_1r_2\cos\varphi}\right]r_1r_2dr_1dr_2d\varphi\\&=\frac{2}{\pi R_0^4}\int_0^{R_0}\int_0^{R_0}\exp\left[-(r_1^2+r_2^2)\right]\int_0^{2\pi}\exp\left[-(2r_1r_2\cos\varphi)\right]d\varphi r_1r_2dr_1dr_2.\end{aligned} \tag{36}$$

Proceeding as above, we first apply (27):

$$\beta^2(R_0)=\frac{4}{R_0^4}\int_0^{R_0}\int_0^{R_0}\exp\left[-(r_1^2+r_2^2)\right]I_0(2r_1r_2)r_1r_2dr_1dr_2 \tag{37}$$

and then (29), to obtain

$$\beta^2(R_0)=\frac{4}{R_0^4}\sum_{k=0}^{\infty}\frac{1}{(k!)^2}\int_0^{R_0}\int_0^{R_0}\exp\left(-(r_1^2+r_2^2)\right)(r_1r_2)^{2k}dr_1dr_2. \tag{38}$$

The double integral in (38) is simply a product of two equal intergals or, in other words, the square of an integral of type (32). We thus obtain, using the notation (31)

$$\beta^2(R_0) = \frac{1}{c^2} \sum_{k=0}^{\infty} \left[1 - \exp(-c) \sum_{i=0}^{k} \frac{c^i}{i!} \right]^2 \qquad (39)$$

Once again, it is very easy to perform computations with the equation (39), using a simple code.

As shown in Section 2, it is not the cross-correlation function γ between the area-averaged and point values but the normalized cross-covariance function ζ that directly enters all equations dealing with the averaging accuracy estimates and with the optimal averaging. If, nevertheless, we are interested in the cross-correlation function as well, it may be easily computed by equation (18), which in our case takes the form

$$\gamma(R, R_0) = \frac{\zeta(R, R_0)}{\beta(R_0)}. \qquad (40)$$

As we use relative distances (those divided by the correlation radius) and normalized variances and covariances, the variance of area-mean β^2 is, in our toy-example, function of only one variable, the (relative) radius of the area R_0 (Fig. 1). It decreases, of course, with increasing R_0, but rather slowly, much slower than the point correlation μ does with increasing distance r. As to the cross-correlation γ (40), it is a function of two variables, R and R_0 (Fig. 2). For a very small area, it is close to $\mu(R)$. Two effects take place as R_0 increases: the "zero intercept" $\gamma(0, R_0)$, that is, the correlation between circle-averaged and central values, decreases, and the cross-correlation radius increases, so that the larger R_0 the flatter is the curve $\gamma(R)$.

Fig. 1. Relative standard deviation (STD) of area-averaged values

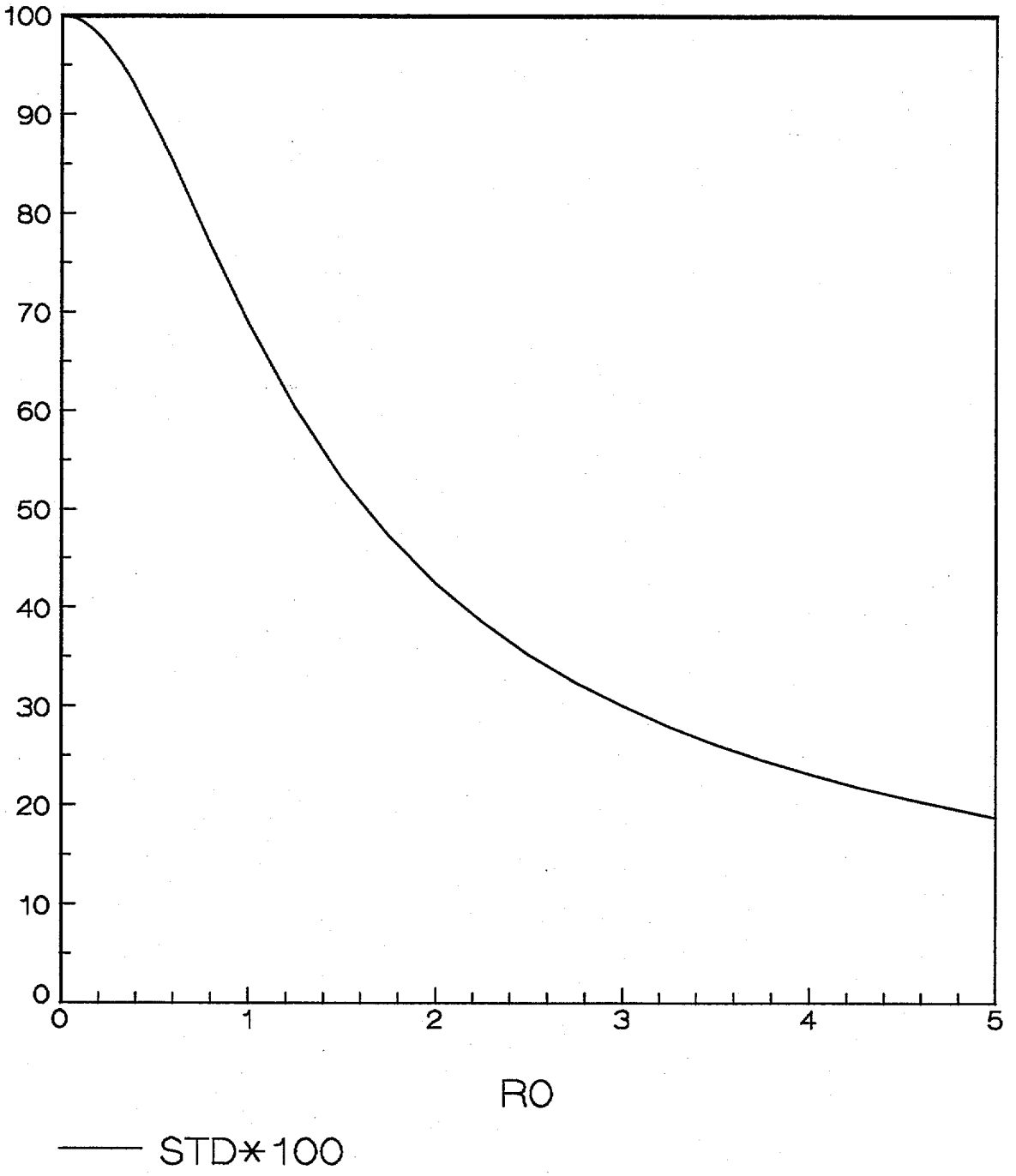
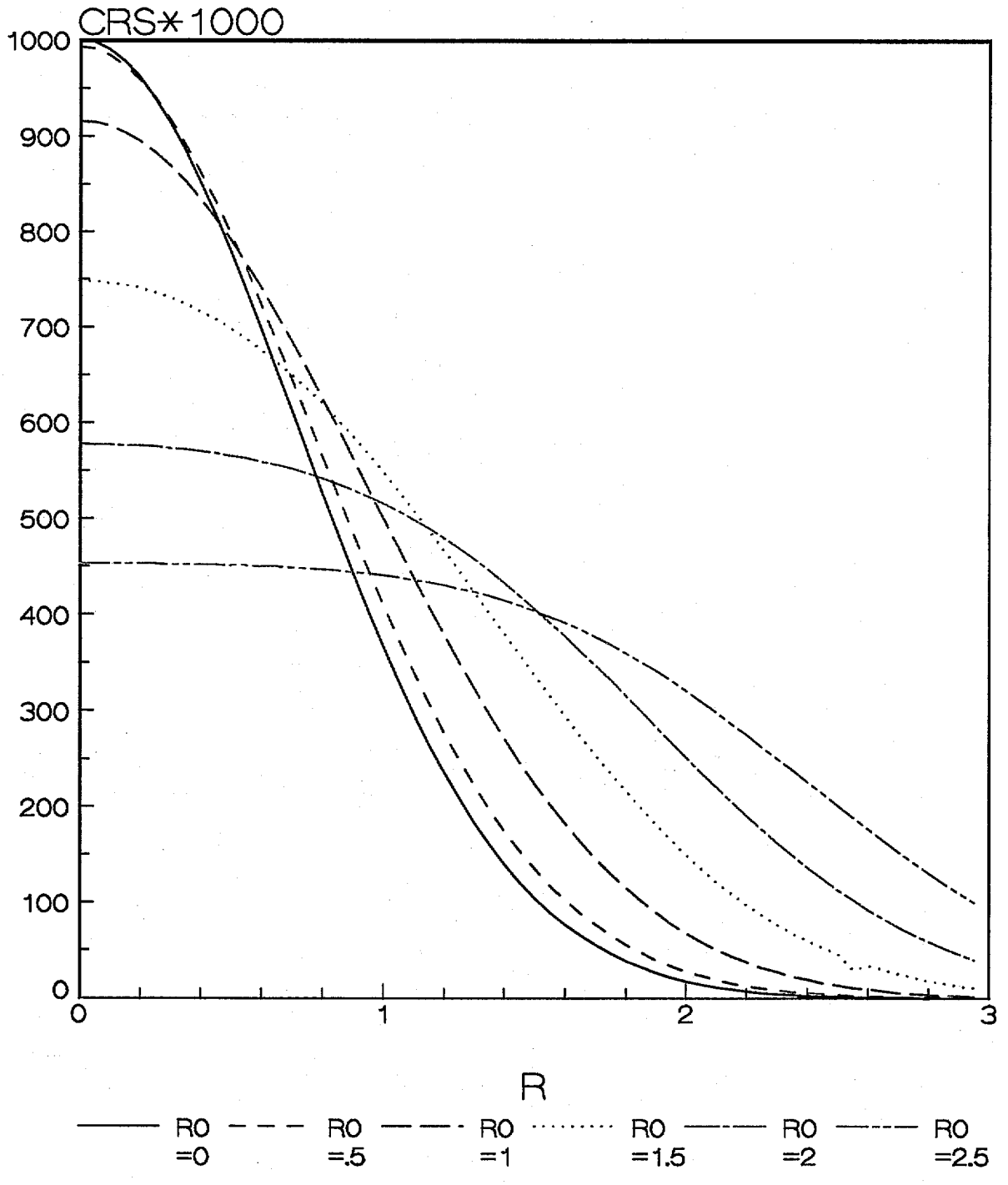


Fig. 2. Cross-correlation (CRS)
between point and area-averaged values



4. REPRESENTATIVENESS OF A POINT OBSERVATION FOR AN AREA-MEAN VALUE.

This section deals with the simplest situation, when only one point observation is used to represent an area-averaged value. The usual means of doing so is just to assume that the area-mean value is equal to observed point value. In order to estimate the RMS error ε of this assumption, one has to put n=1 and w=1 in expression (16):

$$\varepsilon^2 = 1 + \eta^2 - 2\zeta + \beta^2 \tag{41}$$

One may instead apply the optimal averaging, using correspondingly simplified equations (19)

$$\left(1+\eta^2\right)w = \zeta \tag{42}$$

and (20)

$$\varepsilon^2 = \beta^2 - \zeta w,$$

which gives

$$\varepsilon^2 = \beta^2 - \zeta^2 / \left(1+\eta^2\right). \tag{43}$$

The accuracy of the area-mean representation by a point value depends on the area size, on position of the observation point inside (or outside) the area and on the accuracy of the observation.

Two values of the relative observation error variance, η^2 (denoted ETA2 in all figures), η^2 = .05 and η^2 = .5, will be mostly used throughout this paper. The first value

is often used for deviations of a "good" meteorological parameter, like temperature or isobaric height, from climatological norm. According to it, the RMS observation error is about 22% of the RMS anomaly. If the forecast first guess is used, instead of climatology, as the background field, then a substantially larger relative observation error variance should be used, because the variance D^2 of deviations from the forecast first guess is about ten times smaller than that of anomalies. The value $\eta^2 = .5$ used to represent this case corresponds to the RMS observation error of about 70% of the RMS deviation from the forecast first guess.

As to the two other influencing parameters, the radius R_o of the averaging area and the distance *d* of the observation point from its center, two kinds of computations using Eqs (41)-(43) were performed, one with d = 0 (representation from the center) and varying R_o, another with some fixed R_o values and varying *d*. Two simple codes merged with those for ζ and β computation were used in the course of these computations. Figs 3 - 6 present some results of these computations. "Opt" in these figures refers to the optimal averaging, "Arith" to assuming the averaged value to coincide with the observed one, "Abs" to normalized RMS averaging errors and "Rel" to relative RMS averaging errors as they were defined in Section 2.

As may be seen from Fig. 3, presenting the case $\eta^2 = .05$ (climatology as the background field), an observed anomaly at the area center represents the area-averaged value fairly well as long as the area radius remains small, not exceeding, say, 30% of the correlation radius. For such areas, it does not make much difference which method of averaging to apply and whether to consider absolute or relative RMS averaging error, because this error is almost entirely determined by the RMS random error of the point observation and because the optimal averaging weight is close to 1. The latter also means

that the result of the optimal averaging depends almost entirely on the observed point value; the climatological field makes very little influence on it.

The situation becomes quite different when the area radius is close to, or even exceeds, the correlation radius. The optimal averaging weight decreases with increasing area size, so that the area mean value of the parameter itself becomes less influenced by observed point value and more by the climatological field. Correspondingly, the advantage of the optimal averaging over "arithmetic" one becomes more and more pronounced as the area size increases.

Along with this regularity, the differences between absolute and relative RMS averaging errors also increase with increasing area size. If, for example, the area radius is equal to the correlation radius (i.e., if $R_0 = 1$), then the relative averaging error is about 75% while the "absolute" error is about 50%.

Fig. 4 presents results of similar computations with $\eta^2 = .5$ modeling the situation when the forecast first guess is used, instead of climatology, as a background field. Although looking quite analogous to those on Fig. 3, the curves on Fig. 4 demonstrate essential difference between the application of forecast background and of climatological one. The optimal averaging weight is substantially less than 1 even for very small areas, indicating that the full area-mean value is strongly influenced by the first-guess field. Consequently, the accuracy of optimal averaging, or optimal representation, is in this case substantially higher, even for small areas, than that of assuming the area-mean value to be equal to observed point value. This means that, while an "automatic" representation of area-mean value of an anomaly by observed one is acceptable for small areas, it is highly desirable to use, instead, the optimal representation when dealing with deviations from the forecast first guess.

Strictly speaking, this conclusion, as well as many others, is valid only if the underlying statistics exactly reflect the reality, which certainly is never the case. It may be shown, however, that the differences between assumed and real statistical structure have comparatively small influence on the results and accuracy estimates of the optimal averaging. This topic is considered in some detail in Section 6 of this Office Note.

Comparing Figs 3 and 4, one can see that differences between corresponding curves on them are larger for smaller area sizes and smaller for larger areas. This reflects the fact that the relative RMS observation error plays major role for averaging over small areas, while for large areas it is the statistical structure of the parameter in question.

It is necessary to take into account, however, that RMS "absolute" errors on Figs 3 and 4 are those normalized on the RMS deviations from the background fields. If we interpret the case with $\eta^2 = .05$ as typical for deviations from climate (anomalies) and that with $\eta^2 = .5$ as typical for deviations of the same parameter from the forecast first guess (increments), then we actually assume that the variance of anomalies is ten times higher than that of increments, which is a good estimate for such parameters as height, temperature and wind. This means that in order to be able to compare the absolute values of RMS averaging errors for anomalies and for increments, one needs first to perform a kind of "renormalization", e.g., to divide the normalized absolute errors for increments by $\sqrt{10} = 3.16$, retaining those for anomalies as they are. It is easy to see then from Figs. 3 and 4 that averaging of increments promises substantially smaller RMS errors even for small areas, than those achievable when averaging anomalies, and this difference increases for larger areas. As may be seen from results presented in the next section, the same is true when data not at one but at several points are used for the area-averaging.

One may also mention on both Figs 3 and 4 that the relative RMS error of the "automatic" representation of the area-mean value by observed one grows very fast with increasing area size. When R_0 exceeds 1, i.e., when the radius of the area is larger than the correlation radius, then this relative error also exceeds about 1, and it becomes much larger for larger areas. This shows that the simple replacement of the area-mean value by the point value makes no sense for such areas, while the optimal representation may still provide some information about the area-averaged value.

All the estimates above are for the most favorable position of the observation point, i.e., in the center of the area. As may be seen from Figs 5 and 6, the averaging accuracy decreases rather fast with increasing distance *d* of the observation from the area center. When dealing with the representativeness of a single-point observation, one may just assume that such observation should represent the value averaged over some small area around the observation point. If, however, we are interested in recognizing a small signal in the presence of strong noise, as is the case in the problem of climate change detection on the background of natural weather variability, then it is desirable to average over large areas, because, as illustrated by Fig. 1, the larger the area, the smaller is the natural variability of the area-mean value. It is clear from the estimates presented in this section that observation data not at one but at many points should be used in attempts to achieve this aim. This general case of area averaging will be considered in some detail in the next section of this Office Note.

Fig. 3. Representation by central point; ETA2=.05

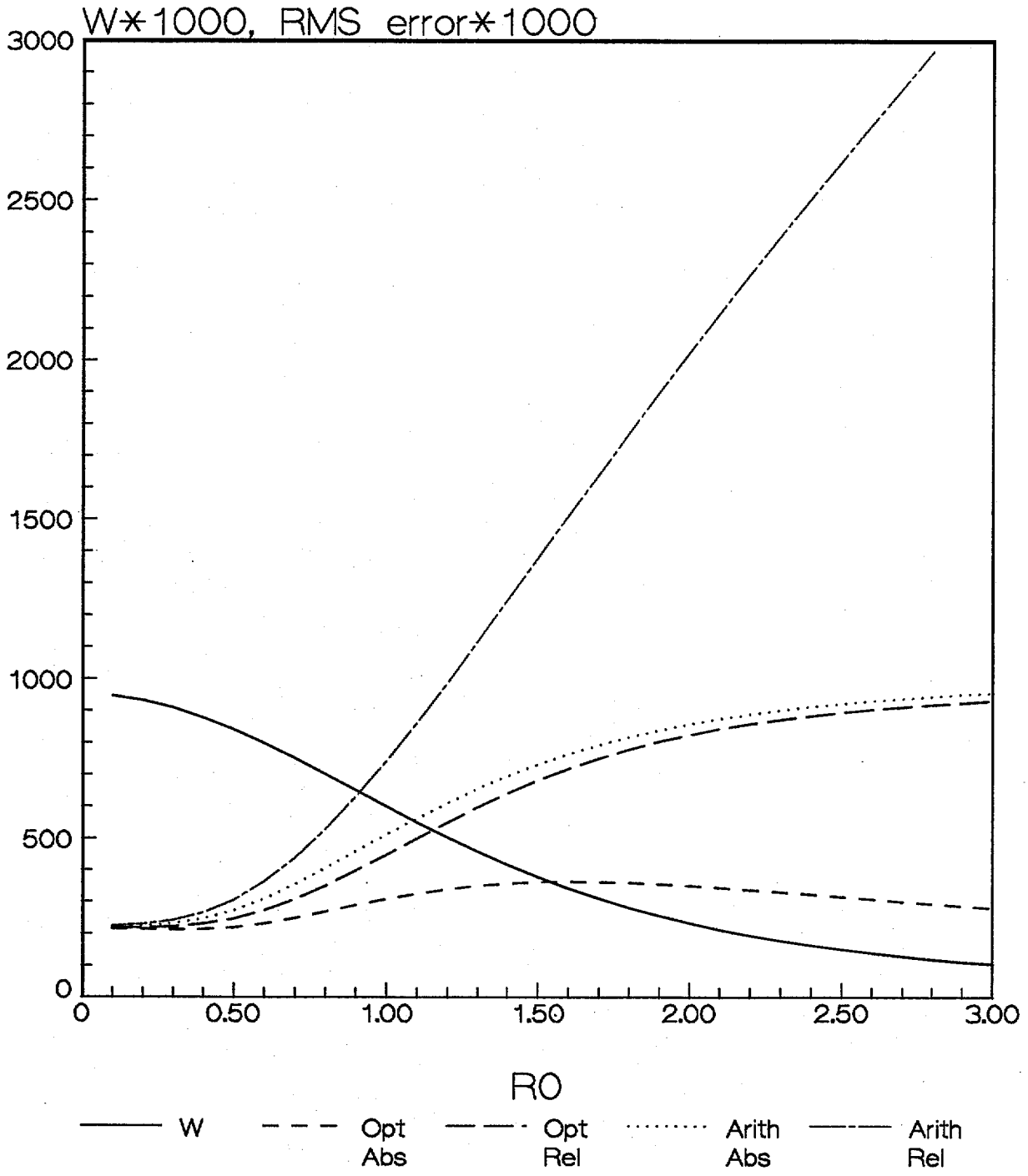


Fig. 4. Representation by central point; ETA2=.5

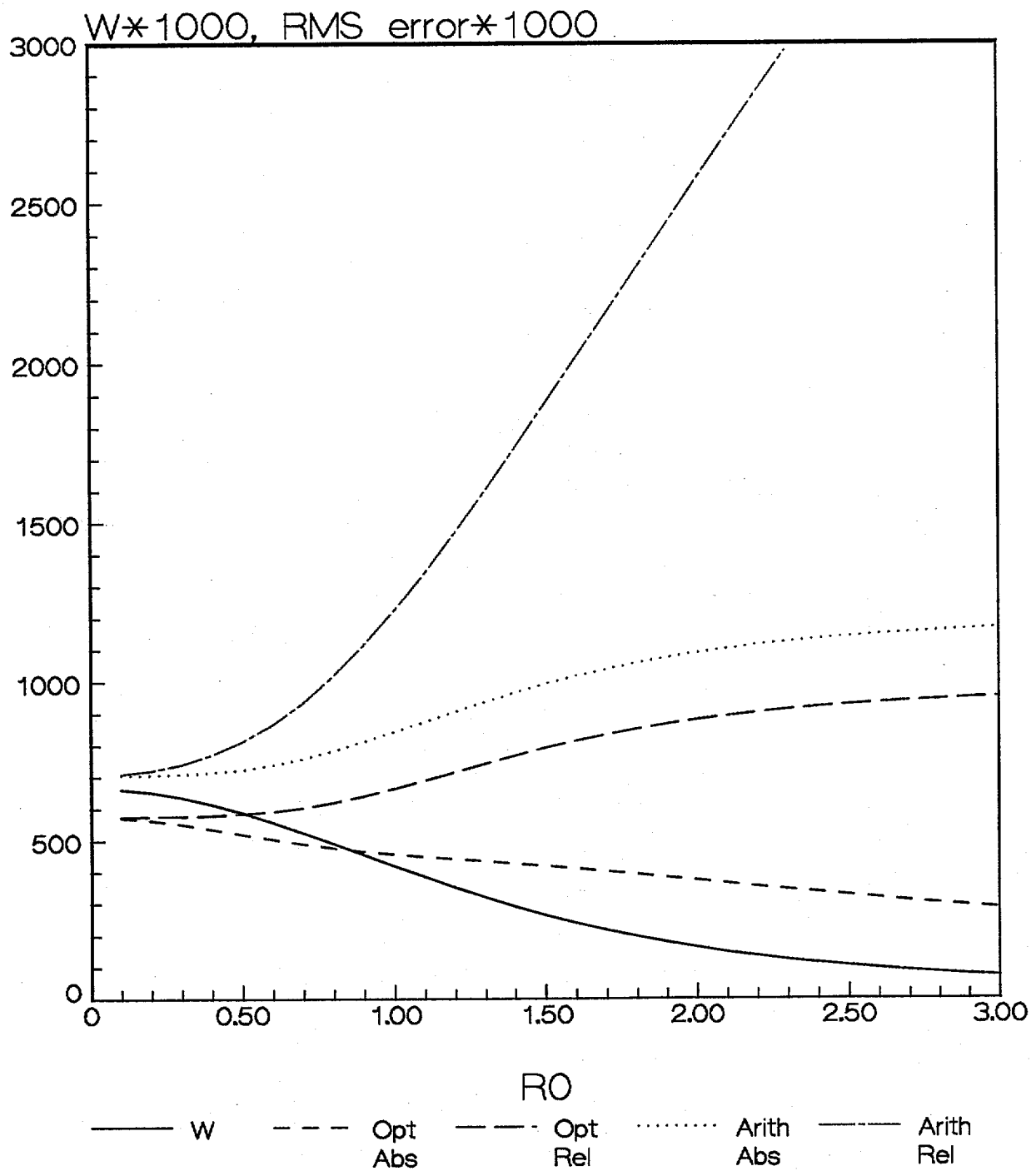


Fig. 5. Representation by various points; R0=1; ETA2=.05

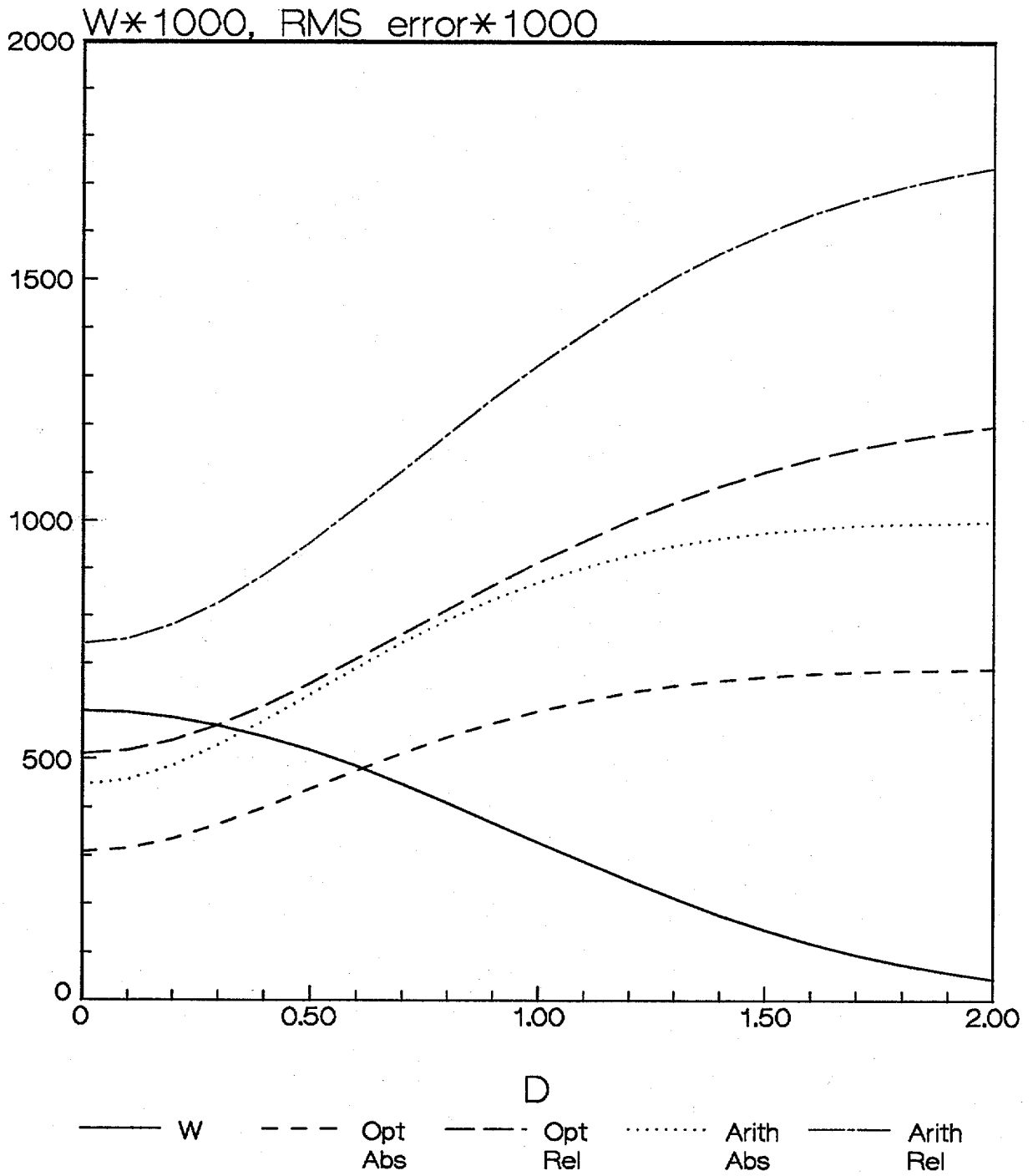
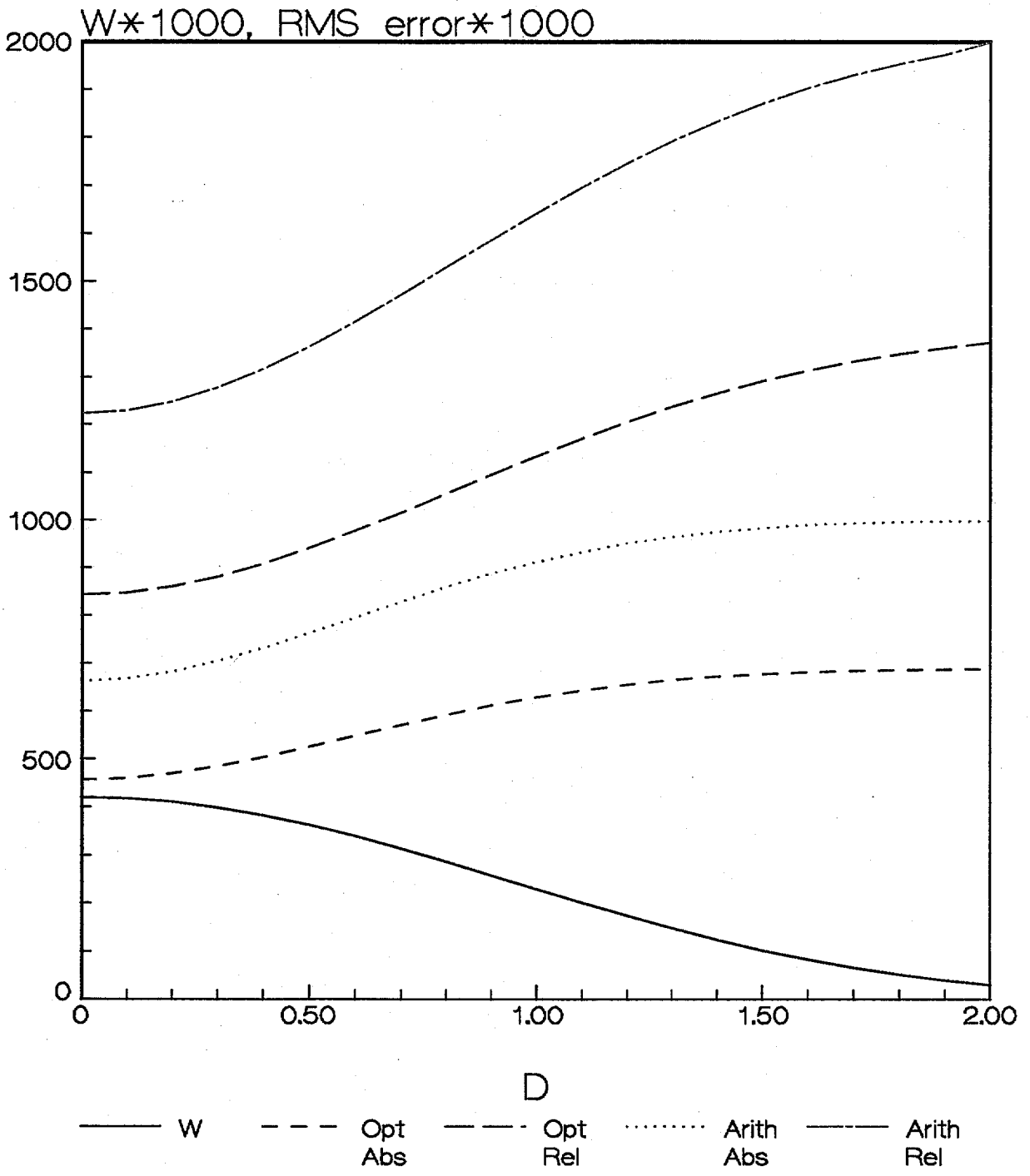


Fig. 6. Representation by various points; R0=1; ETA2=.5



5. ESTIMATION OF AREA-AVERAGED VALUES FROM DATA AT SEVERAL POINTS.

When observed data at several points are used to estimate an area-averaged value, then a variety of factors influence the result of such averaging and its accuracy, and it is practically impossible - and not desirable - to analyse this multi-factor influence in all its complexity. One should try instead to diminish the scope of consideration by using some simplifying assumptions, as well as by analysing the influence of each factor more or less independently of that of other factors.

Applying this approach, we assume, first of all, a rather idealized distribution of observation points: they are supposed to form a rectangular grid (see Fig. 7). This assumption is actually the third simplification leading to our "toy example" (in addition to assumptions that the averaging area is a circle and that the auto-correlation function of point values is Gaussian). This assumption enables us to analyse the influence of various factors in a much simpler way.

To further facilitate this analysis, it is desirable to specify some, so to say, "basic state", i.e., to choose some configuration as a basis for comparisons. Based on some preliminary computations, we have selected, for such basic state, $R_0 = 2$ (i.e., the radius of the averaging area equal to twice the correlation radius), 25 observations covering the 2x2 square centered in the averaging circle, and absence of correlation between the random observation errors. As to the relative variance of the observation errors, two above-mentioned values, $\eta^2 = .05$ (for anomalies) and $\eta^2 = .5$ (for deviations from forecast) will be considered.

The influence of the number of "stations" NP on the relative RMS error of optimal and arithmetic averaging, as well as on the sum of optimal averaging weights, is shown on Figs 8 and 9. One can see on these graphs that the weight sum quickly increases and the optimal averaging error quickly decreases with increasing NP for small NPs, remaining

almost constant for larger NP values (a minor "irregularity" between NP = 9 and 16 on Fig. 8 is caused by the fact that 9 points include that in the center of the area, while 16 do not). As to the arithmetic averaging, its RMS error also decreases for small NPs, but then it begins to increase. This effect is particularly pronounced in Fig. 8, indicating that using larger number of points in the course of arithmetically averaging anomalies over an area may be even harmful.

In this respect, it is important to stress the following point. The so-called "square root law", stating that the RMS error in an average of N data is proportional to $N^{-.5}$, is often used to estimate the gain in accuracy caused by the averaging. This law is, however, applicable only to the averaging of *independent* data. As to the values of a meteorological parameter at different points, these values are interdependent, the more so the closer are the points to each other. The influence of this may be seen from the curves "indep" on Figs 8 and 9, reflecting the RMS error behaviour (with respect to its value for NP = 4) according to the square root law. It may be even better demonstrated by Figs 10 and 11, showing the same curves as Figs 8 and 9 (with added points for NP = 1) in a logarithmic scale, so that the square root law is described by a straight line. (There are two such lines of each of Figs. 10 and 11: on, denoted "SQR eta", crosses the NP=1 line at $\varepsilon = \eta$, another, SQR opt1, crosses this line at $\varepsilon(1)$).

The general conclusion from these examples is rather optimistic: they show that a gain in accuracy achievable by using data at more than one observation point to estimate area-mean value over a comparatively large area is quite high, particularly if the optimal averaging is applied. At the same time, a naive assumption, that this gain may continue infinitely when the observation density increases, would be, of course, wrong.

In fact, results like those may be used in order to design a rational station network density for presentation of area-mean values, just like the results of analogous

computations with the optimal interpolation instead of averaging are used to evaluate the rational network density for representation of point values between the observation points.

One can mention in this respect that, from a qualitative point of view, the described regularities with averaging do not differ from those valid for interpolation. Quantitatively, however, they differ very much. The main fact is that the RMS absolute (normalized) optimal averaging error is much less, under other equal conditions, than the RMS optimal interpolation error. One can also see that the pattern of optimal averaging weights is quite different from that of optimal interpolation weights. This difference may be illustrated by weight patterns on Figs 12 and 13. While the interpolation weight is maximal for the central point and decreases fast towards the area boundary, the pattern of averaging weights is much more homogeneous.

A different kind of functional dependence is presented in Fig. 14. The number of observation points was held constant (NP = 25) for this series of computations, but the distance between neighboring points and, consequently, the size of area covered by these points varied. As may be seen from Fig. 14, there exists (for a fixed NP) an optimal density of such idealized network, such that the RMS averaging errors increase when the density becomes higher or lower than the optimal one. This effect, typical for any averaging, is a consequence of the fact that both limiting cases, XN = 0 and XN = ∞ , are equivalent to the use of only one observation point. It is important to mention, however, that this effect is quite strong for the arithmetic averaging, while, for the optimal averaging, it is almost negligible. This example demonstrates another advantage of the optimal averaging over the arithmetic one, its lower sensitivity to, so to say, external conditions of the averaging.

The same conclusion may be drawn from curves on Figs 15 and 16 reflecting series of computations with a fixed number (25) and disposition (in the 2x2 domain) of the observation points but with varying radius R_0 of the averaging area. Just like in the

previous series (Fig. 14), there exists an optimal R_0 value, and just like in that series, the accuracy of arithmetic averaging markedly decreases when R_0 differs from this optimal value (particularly, when it becomes larger), while the optimal averaging accuracy is much less sensitive to this factor.

The dependence of the area averaging accuracy on the remaining parameter, the (normalized) RMS random observation error η, is, unlike those shown on Figs 14-16, monotonous. As illustrated by Fig. 17, RMS errors of both optimal and linear averaging increase with growing η, while the sum of optimal averaging weights, as well as their variance, decreases.

All estimates above reflect the most favorable position of the observation network, that centered with respect to the averaging area. In other words, it was assumed up to now that the center of the "station" grid coincides with the averaging area center. As may be seen from Figs 18 and 19, the situation becomes much worse when this symmetry is violated. Certainly, these graphs reflect extreme cases when the observation grid covers only one half of the averaging area, but they demonstrate quite convincingly that it is very important to have the set of observation points as much centered with respect to the averaging domain as possible. This means, from a practical point of view, that with a given network of stations, one should select every averaging area in such a way that the corresponding stations cover the area most homogeneously and symmetrically.

The curves on Figs 20 and 21 are based on results of computations already reflected by Figs 15, 16, 18 and 19, but "absolute" RMS averaging errors are presented, instead of relative ones, on Figs 20 and 21, and they have been "renormalized" in order to compare the accuracy of averaging (both arithmetic and optimal), achievable for deviations from climatological first guess (anomalies) and for deviations from forecast first guess (increments). As mentioned in Section 4, the variance of increments is about 10 times smaller than that of anomalies, and this fact is reflected in our computations by

assuming the normalized observational error variance η^2 for increments 10 times larger than that for anomalies. To make the absolute RMS errors for anomalies and increments normalized by the same value (the standard deviation of anomalies) and thus comparable with each other, one has to divide all normalized absolute errors of area averaging for increments (i.e., with $\eta^2 = .5$) by $\sqrt{10} = 3.16$, and this has been done when obtaining results illustrated by Figs 20 and 21. Fig. 20 is, like Figs 15 and 16, for the case of centered network of 25 "stations", while Fig. 21, like Figs 18 and 19, illustrate the case of a network over a half of the domain.

As may be seen from Fig. 20, it makes a large difference which averaging method to apply or whether to average anomalies or increments, except when the disposition of observation points is close to most favorable one (which is practically almost never the case). The less favorable is the averaging situation, the higher is the gain in accuracy due to using optimal averaging instead of arithmetic one and to averaging increments instead of anomalies. These differences are particularly pronounced in the case of the network covering only a half of the domain (Fig. 21). The optimal averaging accuracy for increments may be still acceptable in cases like this, while the use of anomalies instead of increments and the application of arithmetic averaging instead of optimal one both result in substantial decreases of the accuracy.

The last question to be considered in this section is the influence of a spatial correlation between the observation errors on the optimal averaging procedure and on the averaging accuracy. This problem arises, first of all, in connection with the possible application of satellite indirect sounding data in addition to, or even instead of data from stationary observation points, in the computation of area-mean values, as is done routinely in the course of interpolation for objective analysis of meteorological fields. It is well known that the random errors in satellite retrieved data are, unlike those in rawinsonde

information, horizontally correlated, and this correlation of observation errors substantially influences the interpolation procedure, as well as its accuracy.

In order to investigate the influence of random observational error correlation on the area-averaging, it is necessary, first of all, to replace the equation (5), which postulates the absence of error correlation (see Section 2), by a more general equation

$$\overline{\delta_i \delta_j} = \Delta^2 \nu_{ij}, \tag{44}$$

where ν_{ij} is the coefficient of correlation between the observation errors at points i and j. We will assume for the observation errors, as for parameter itself, that their correlation is homogeneous and isotropic, i.e., that the coefficient ν_{ij} depends only on the distance r_{ij} between the points.

Under such assumptions, the equation (16) should be replaced by

$$\varepsilon^2 = \sum_{i=1}^{n}\sum_{k=1}^{n} w_i w_k \left(\mu(r_{ik}) + \eta^2 \nu(r_{ik})\right) - 2\sum_{i=1}^{n} w_i \zeta_i + \beta^2 . \tag{45}$$

Correspondingly, the system (19) should be replaced by

$$\sum_{k=1}^{n} \left(\mu(r_{ik}) + \eta^2 \nu(r_{ik})\right) w_k = \zeta_i \qquad (i = 1, 2, \ldots, n), \tag{46}$$

while the equation (20) remains, formally, unchanged.

Results presented in Figs 22 and 23 were obtained under an additional assumption that the error correlation function may be described by the same Gaussian function as the correlation of parameter itself, but with a different correlation radius, namely

$$\nu(r) = \mu\left(\frac{r}{s}\right), \tag{47}$$

where s, the error correlation scale, is the ratio of the error correlation radius to the parameter correlation radius (recall that the latter has been taken as the unit for measuring

distances). s=0 on Figs 22 and 23 corresponds to non-correlated observational errors, while s=1 relates to the case when the radius of error correlation is as large as that for parameter itself.

Fig. 22 demonstrates that the correlation of observational errors leads to a substantial decrease in the averaging accuracy, when the relative variance of the observation errors is large, and it has practically no impact when the relative variance is as small as .05. This does not mean, however, that the optimal averaging procedure is not influenced by the error correlation. Just like in the case of optimal interpolation, the (positive) correlation of observation errors leads to more uneven, "capricious" pattern of optimal averaging weights, as may be seen from Fig. 23 showing the dependence of minimal and maximal optimal averaging weights from the error correlation scale. This effect is a result of diminished diagonal predominance in the matrix of the system (46) due to increase of non-diagonal elements of this matrix as compared with that for non-correlated errors. Particularly, one can see from (46) and (47), that for s=1 (so that $\nu(r)=\mu(r)$) the matrix is basically the same as in the case of complete absence of random observational errors. This indicates that smoothing of the weight pattern, caused by the presence of non-correlated, or slightly correlated, observational errors, just vanishes if s is equal to 1.

The decrease of diagonal predominance in a matrix inevitably leads to a decrease in the computational stability of the solution of corresponding system of equations. It may even result in violation of the positive definiteness. In order to avoid this danger, it is desirable to use a small number of points with satellite data when performing the optimal averaging over an area, or to use already averaged satellite retrieval data, the so-called superobs.

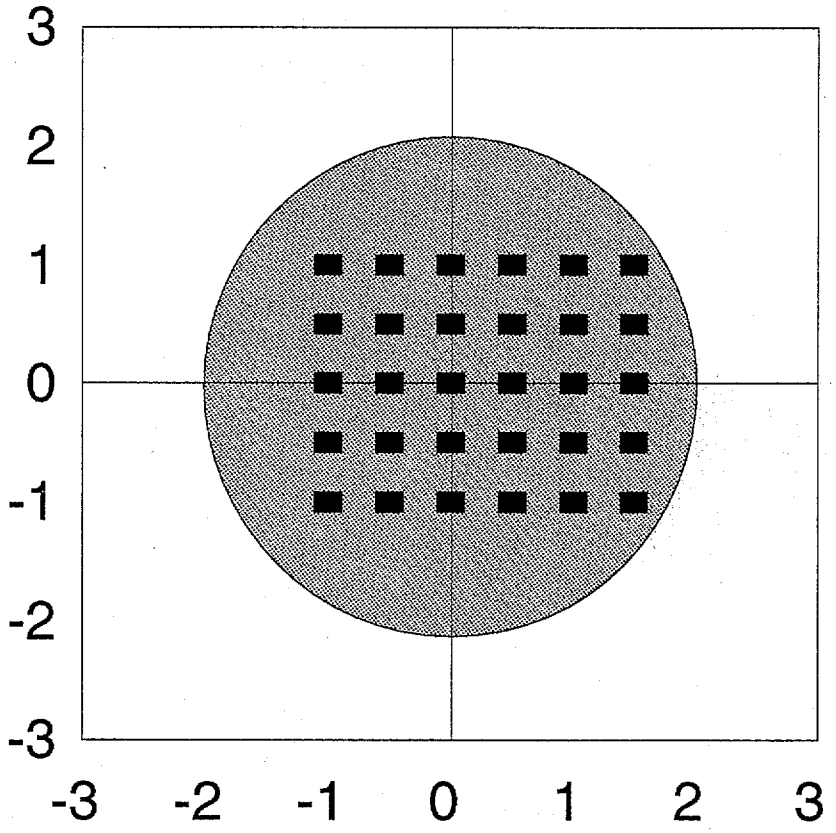


Fig. 7. Averaging area and "observation" points: an example. R0=2; X1=-1; XN=1.5; Y1=-1; NX=6; NY=5; NP=30

Fig. 8. Weight sum and RMS errors
R=2; [−1,1],[−1,1]; ETA2=0.05

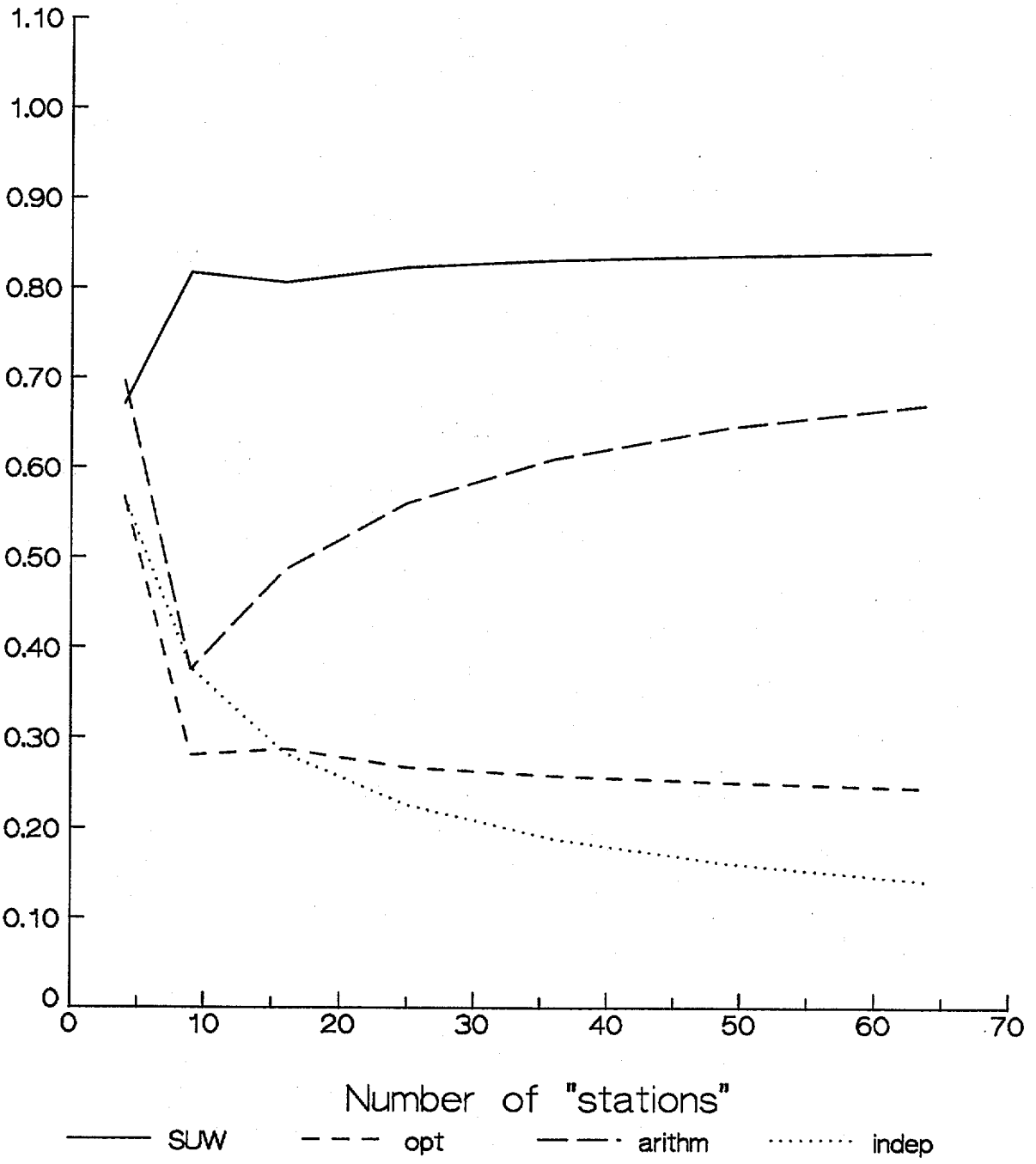


Fig. 9. Weight sum and RMS errors
R=2; [−1,1],[−1,1]; ETA2=0.5

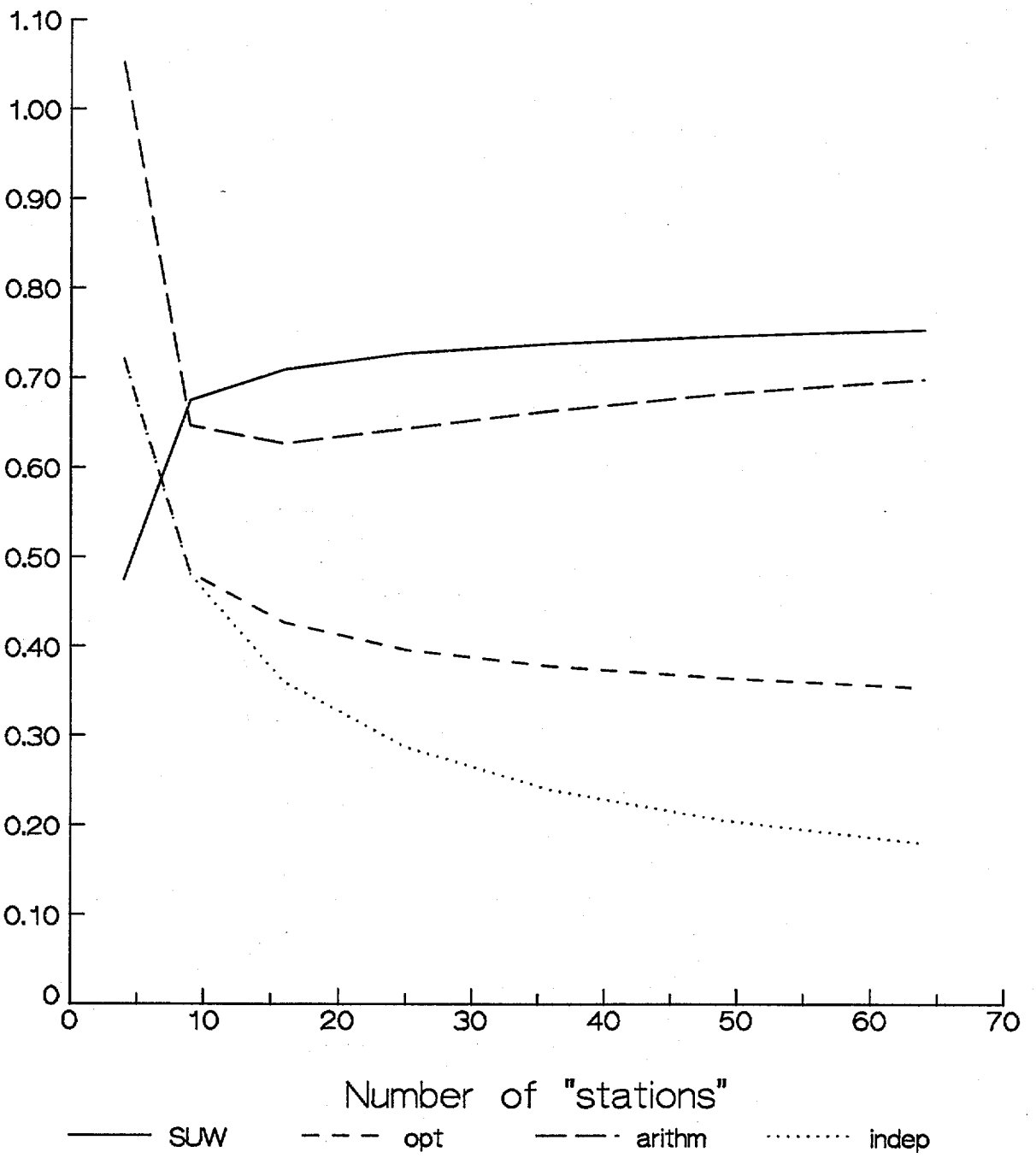


Fig. 10. RMS errors and the "SQR law"
R=2; [−1,1],[−1,1]; ETA2=0.05

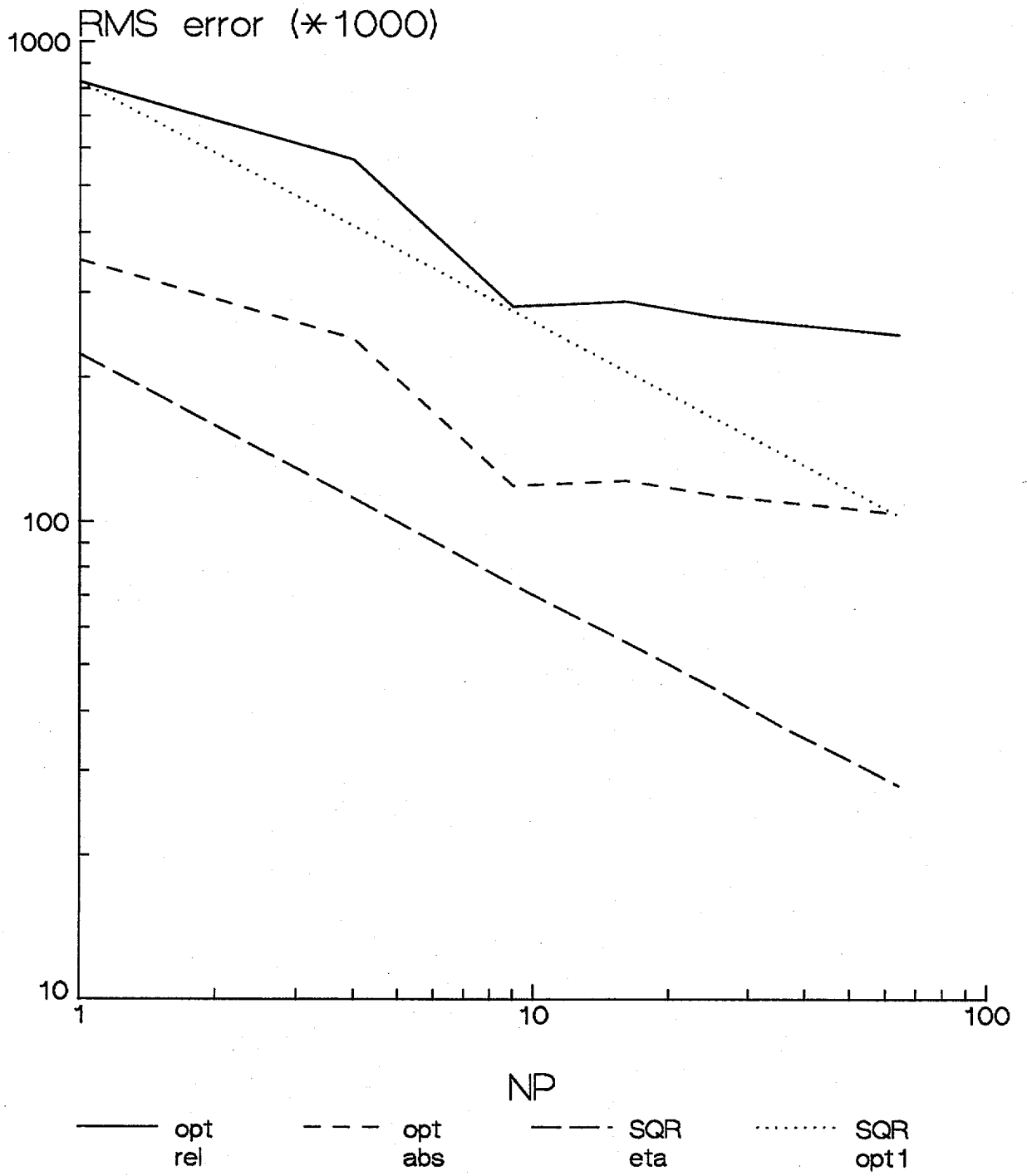


Fig. 11. RMS errors and the "SQR law"
R=2; [−1,1],[−1,1]; ETA2=0.5

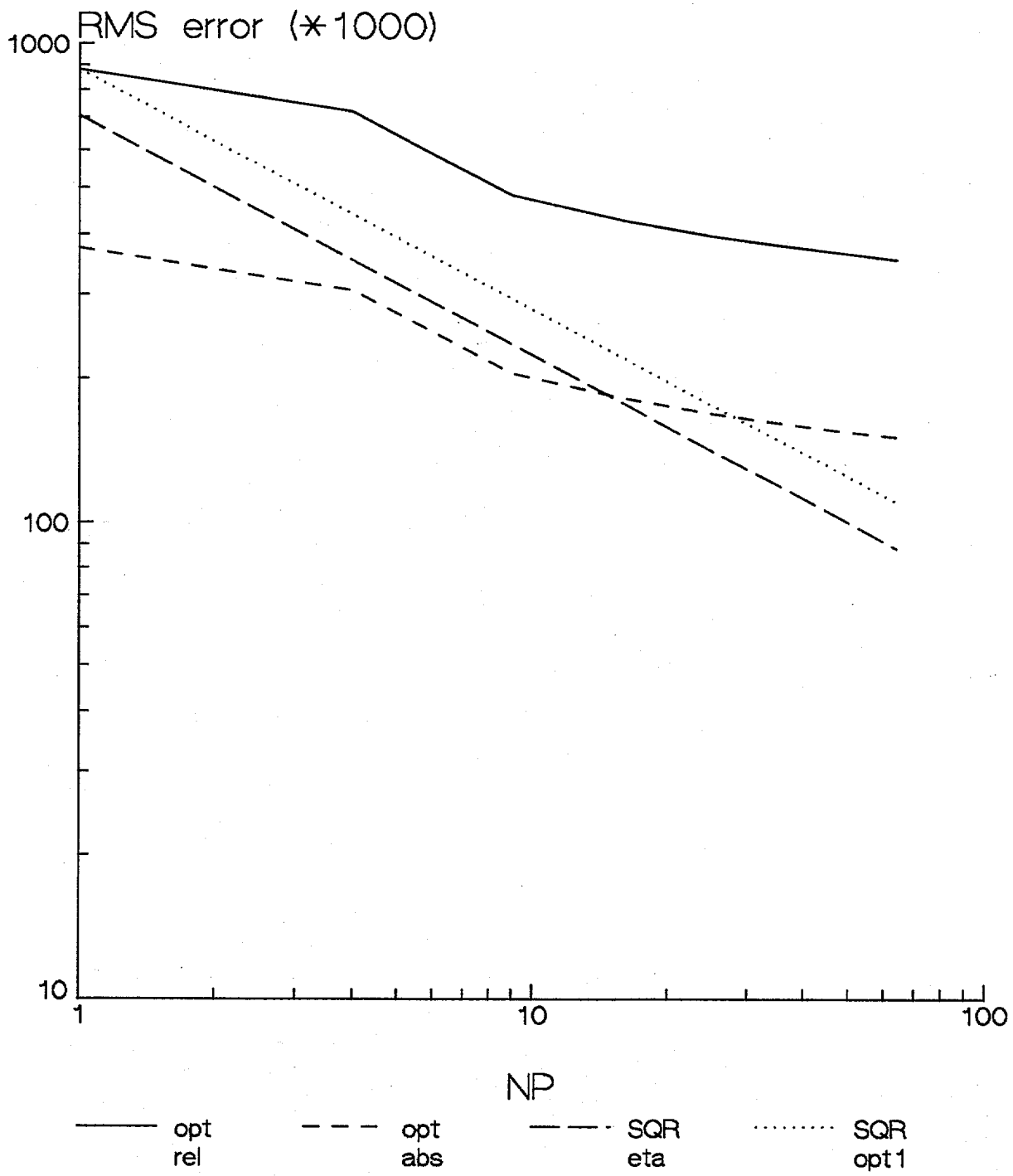


Fig. 12. OPTIMAL AVERAGING versus OPTIMAL INTERPOLATION

[-1,1], [-1,1]. R_0=2. NX=NY=5. ETA2=0.05.

Weight sum and RMS errors

	WSUM	Optimal abs.	Optimal rel.	Arithmetic abs.	Arithmetic rel.
Averaging	0.8231	0.1139	0.2681	0.2383	0.5607
Interpolation	0.9689	0.1519	0.1519	0.6658	0.6658

Weights (*1000)

Averaging					Interpolation				
76	32	77	32	76	16	-35	-25	-35	16
32	-45	12	-45	32	-35	25	180	25	-35
77	12	84	12	77	-25	180	462	180	-25
32	-45	12	-45	32	-35	25	180	25	-35
76	32	77	32	76	16	-35	-25	-35	16

Fig. 13. OPTIMAL AVERAGING versus OPTIMAL INTERPOLATION

[-1,1], [-1,1]. R_o=2. NX=NY=5. ETA2=0.5.

Weight sum and RMS errors

	WSUM	Optimal abs.	Optimal rel.	Arithmetic abs.	Arithmetic rel.
Averaging	0.7272	0.1806	0.3974	0.2735	0.6435
Interpolation	0.9632	0.3629	0.3629	0.6792	0.6792

Weights (*1000)

Averaging					Interpolation				
47	36	37	36	47	-18	-15	-3	-15	-18
36	11	12	11	36	-15	77	150	77	-15
37	12	13	12	37	-3	150	263	150	-3
36	11	12	11	36	-15	77	150	77	-15
47	36	37	36	47	-18	-15	-3	-15	-18

Fig. 14. Weight sum and RMS errors
R=2; XN=YN=−X1=−Y1; NP=25

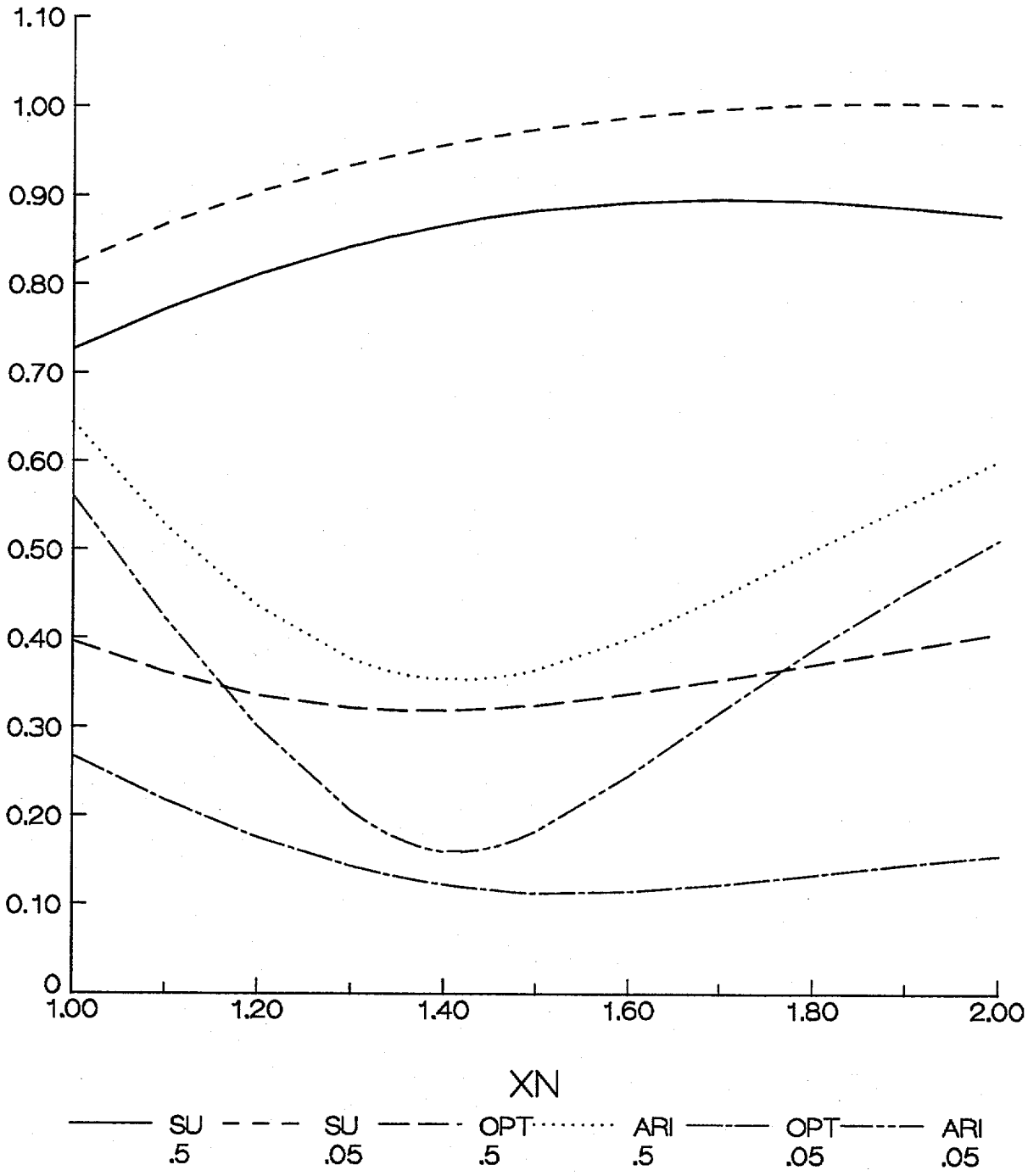


Fig. 15. Averaging over various areas
ETA2=.05; NP=25

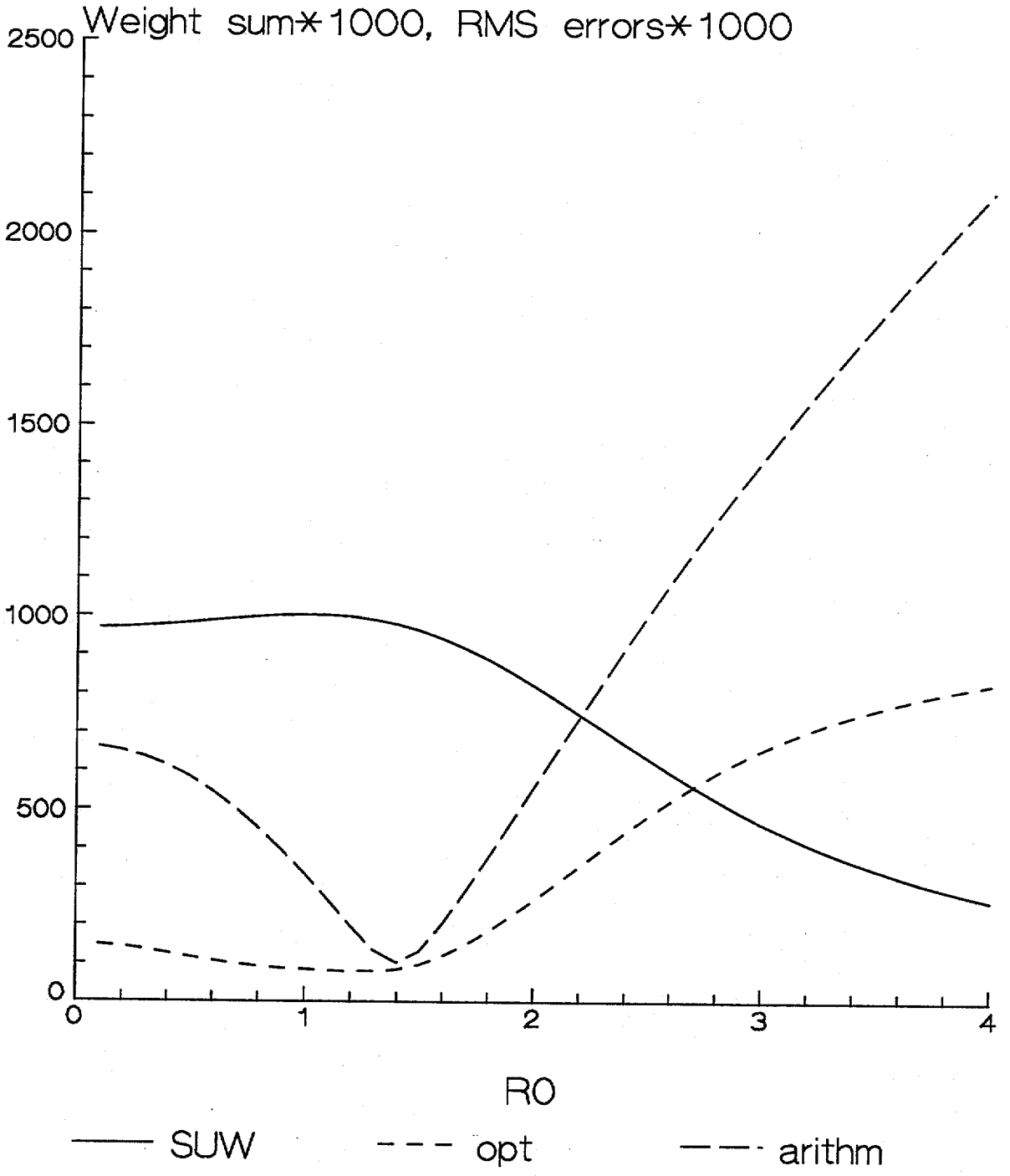


Fig. 16. Averaging over various areas
ETA2=.5; NP=25

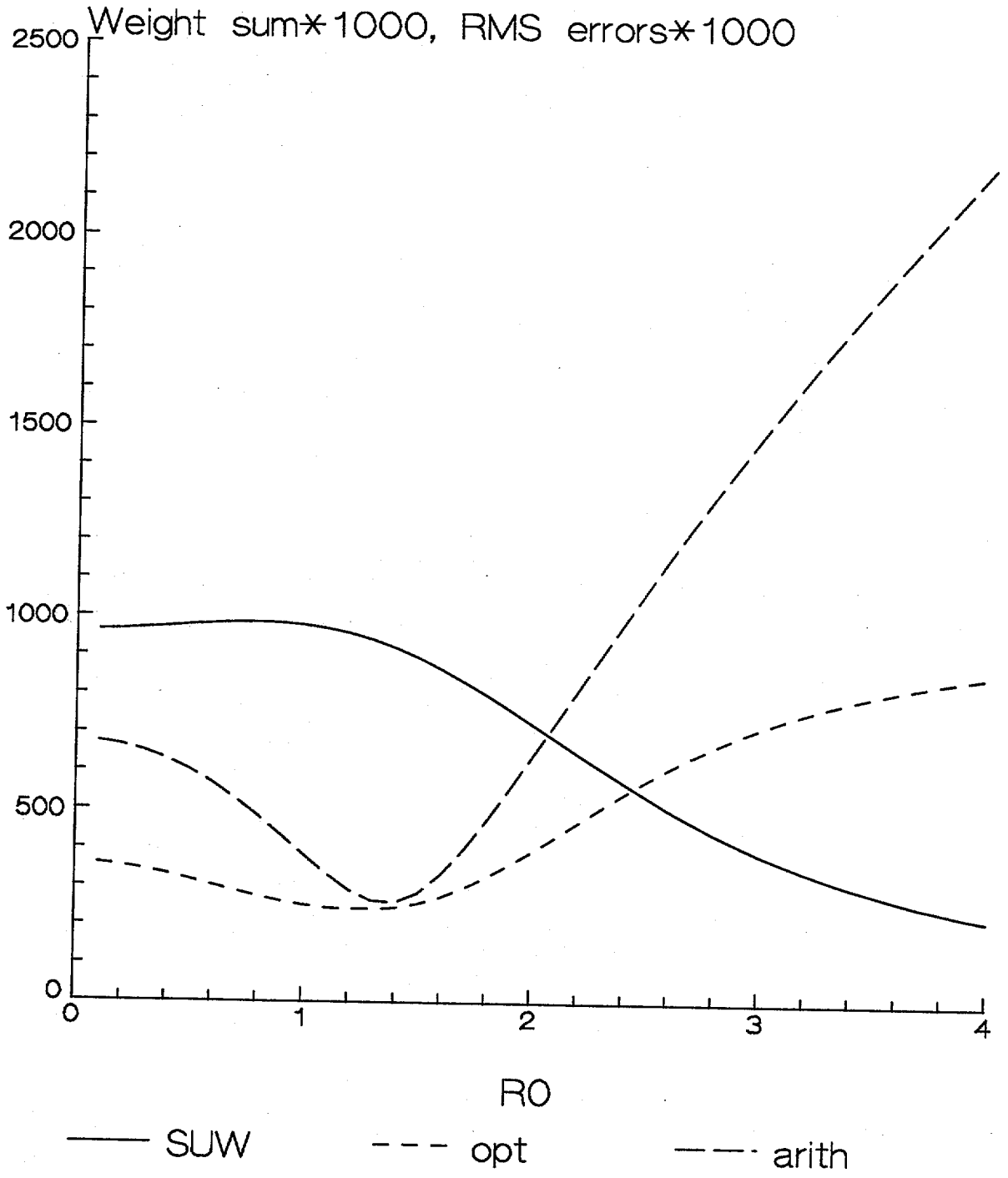


Fig. 17. Area-averaging accuracy
Dependence on eta

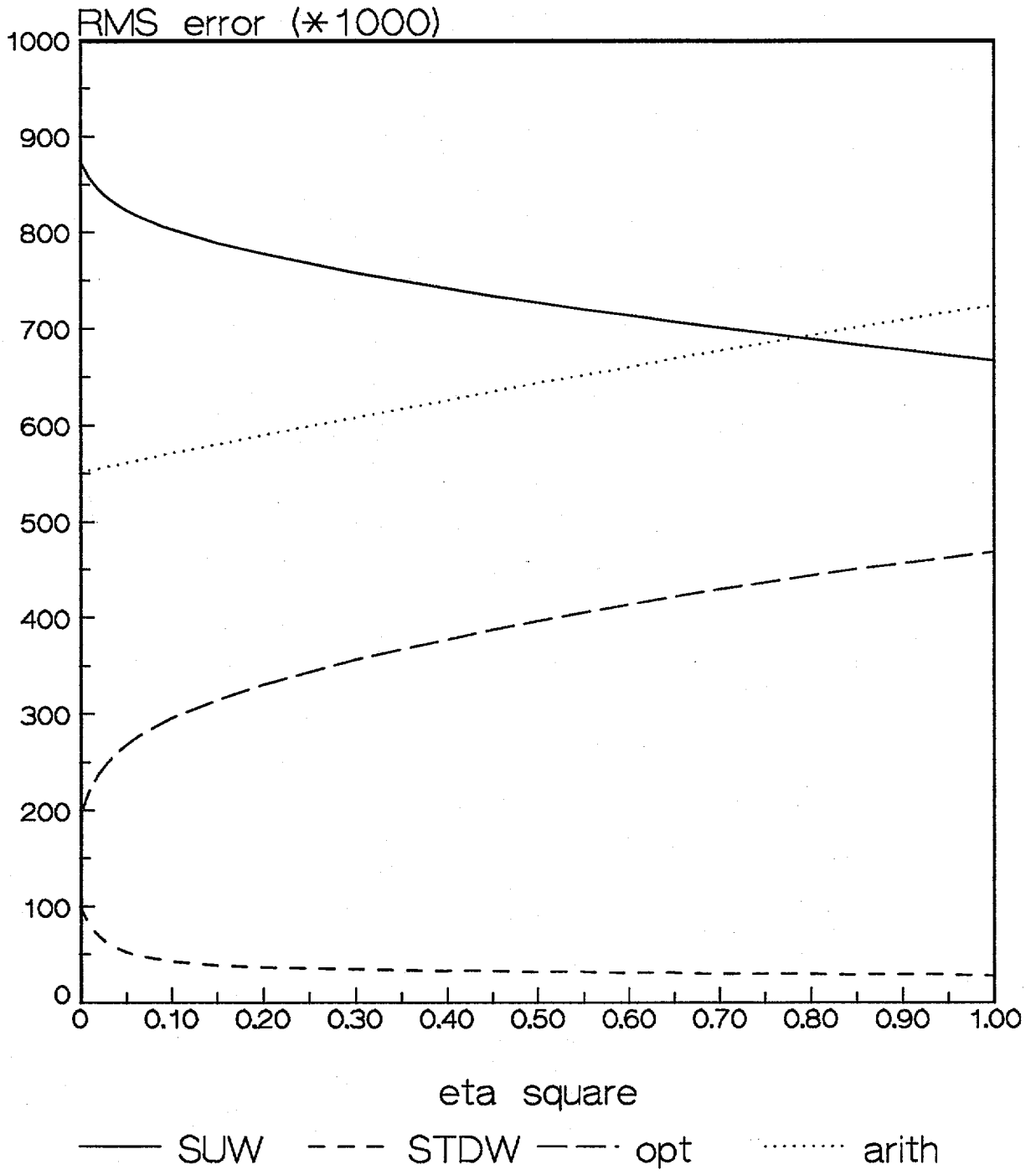


Fig. 18. Averaging with data over a half; ETA2=.05; NP=25

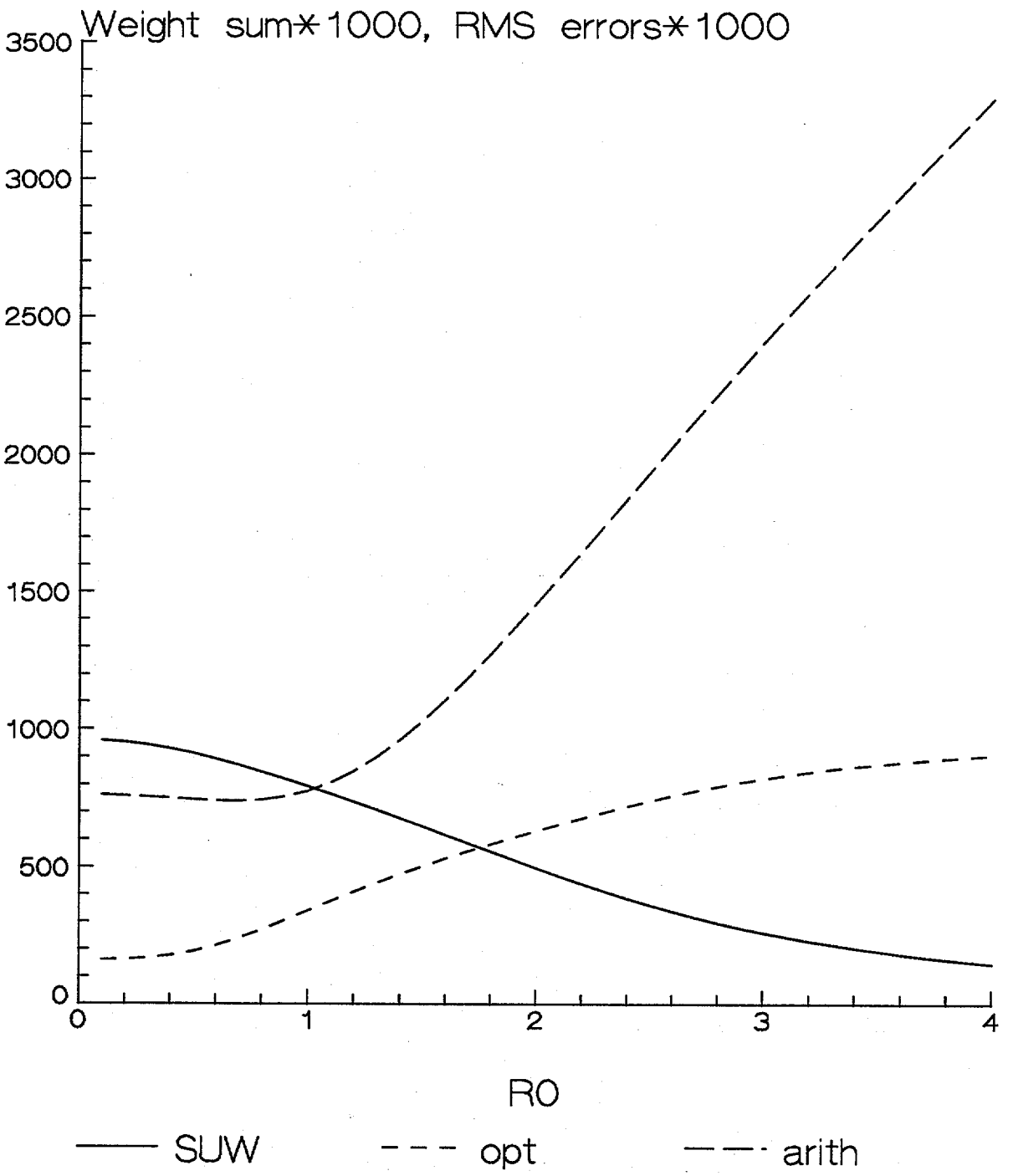


Fig. 19. Averaging with data over a half; ETA2=.5; NP=25

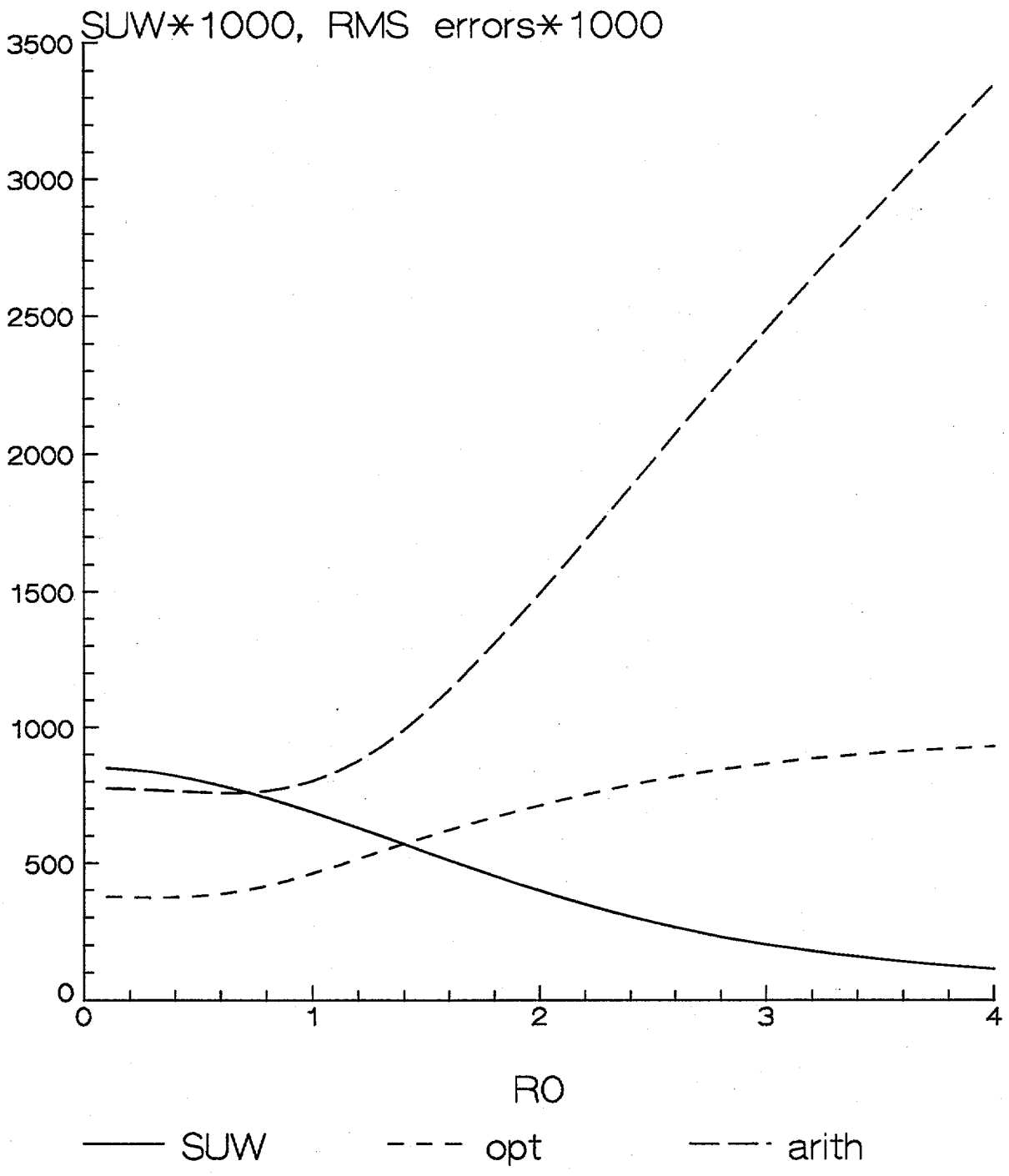


Fig. 20. Renormalized averaging errors; NP=25; [−1,1],[−1,1]

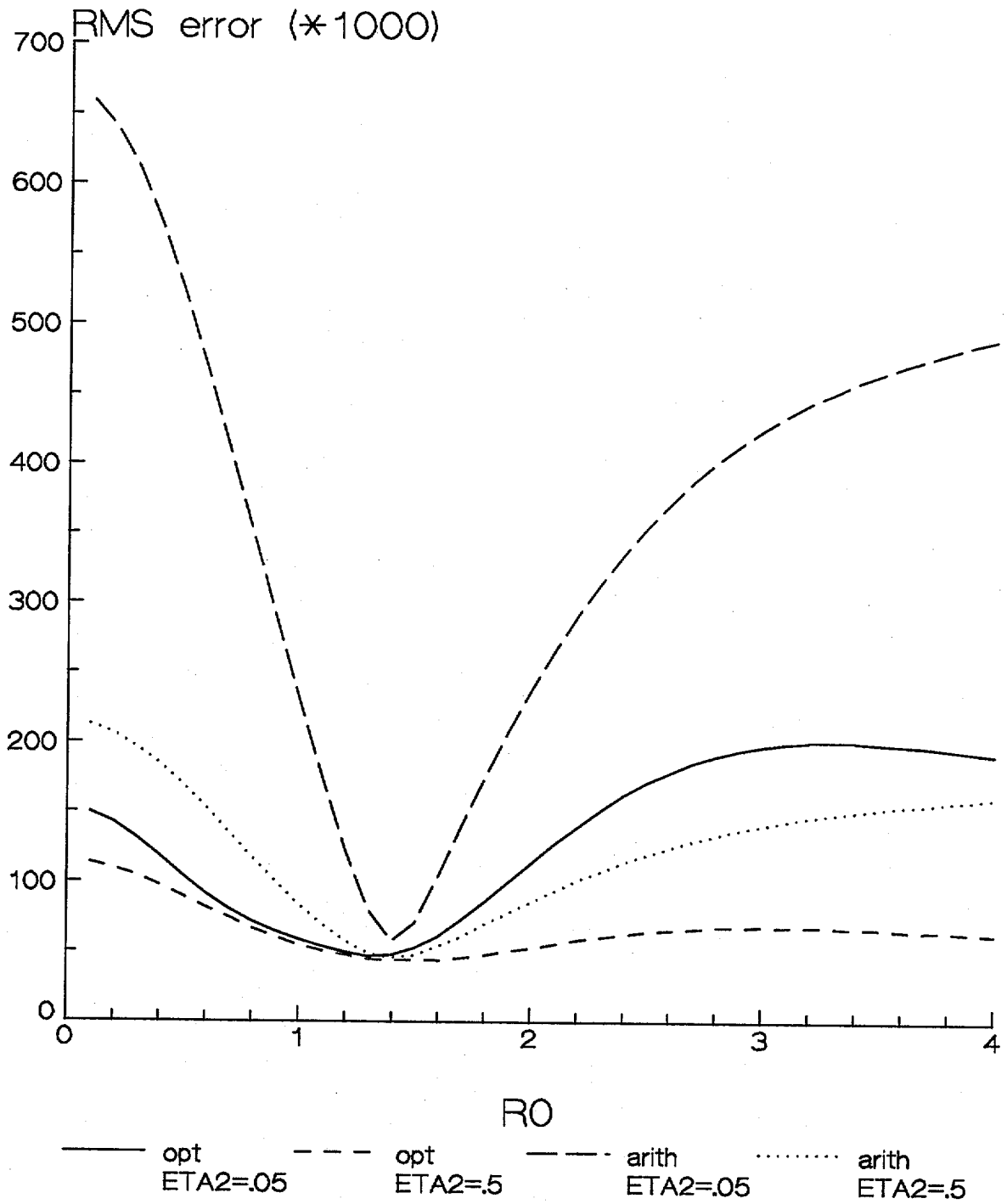


Fig. 21. Renormalized averaging errors; NP=25; [0,1],[−1,1]

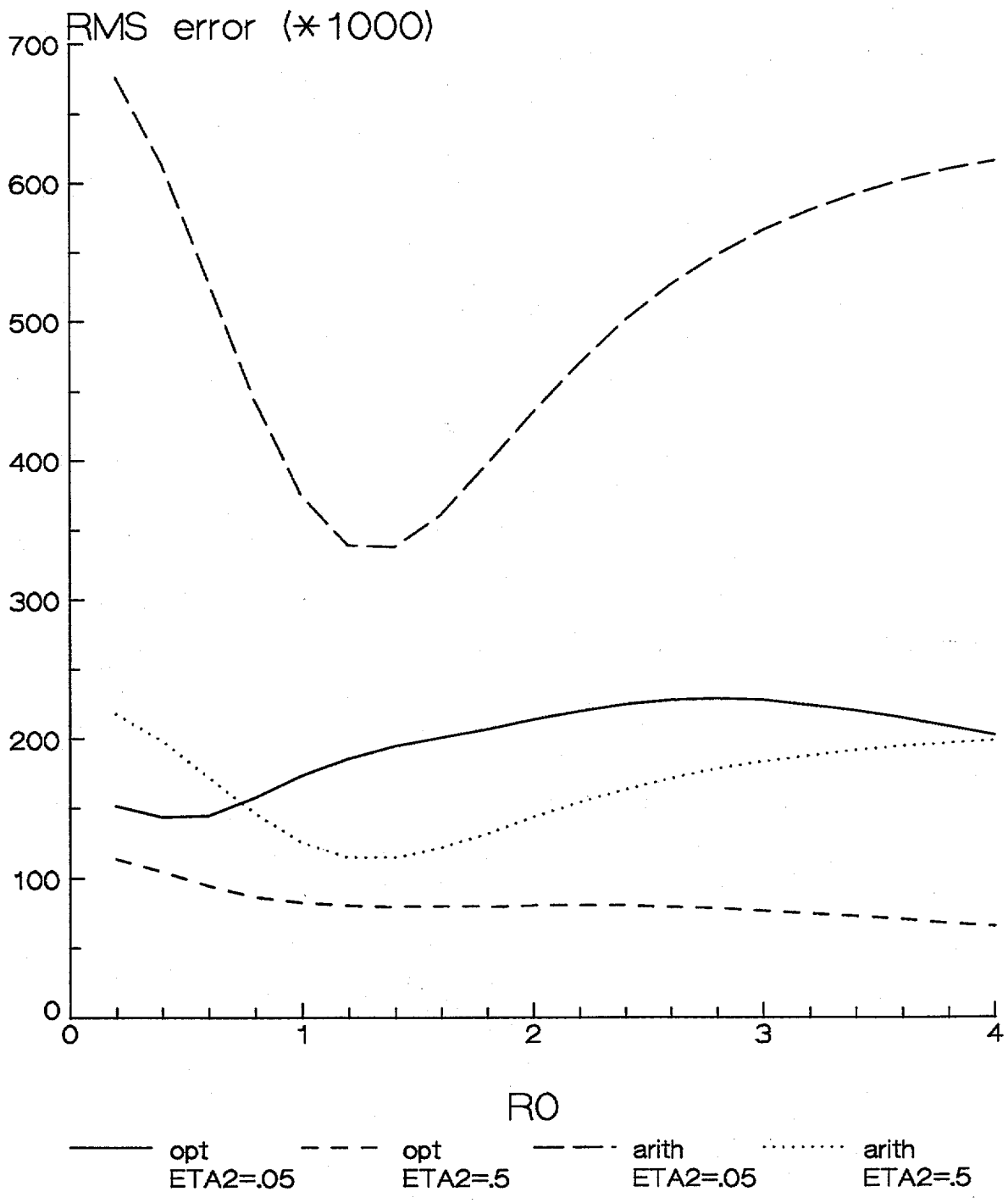


Fig. 22. Averaging of observations
with correlated errors

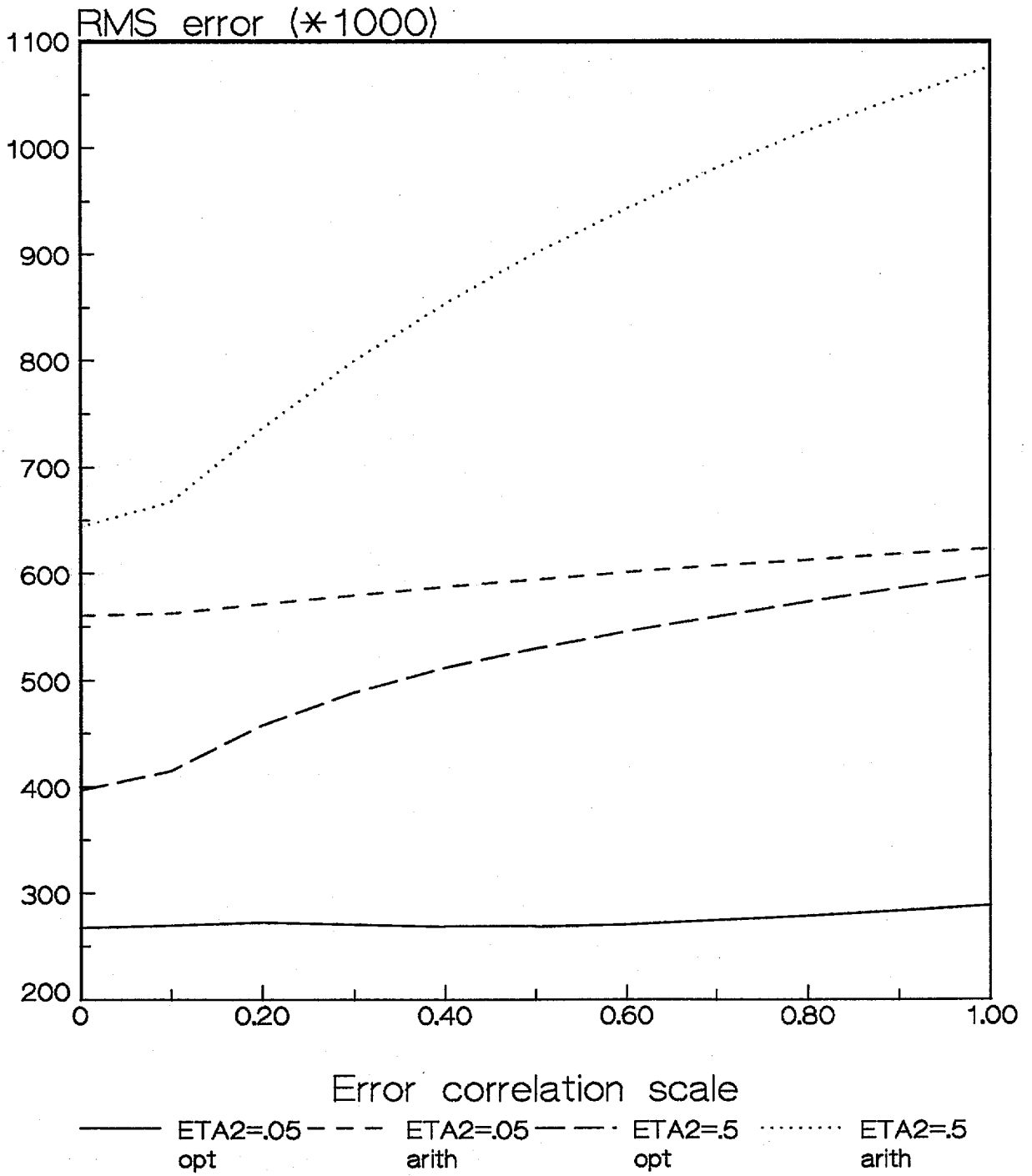
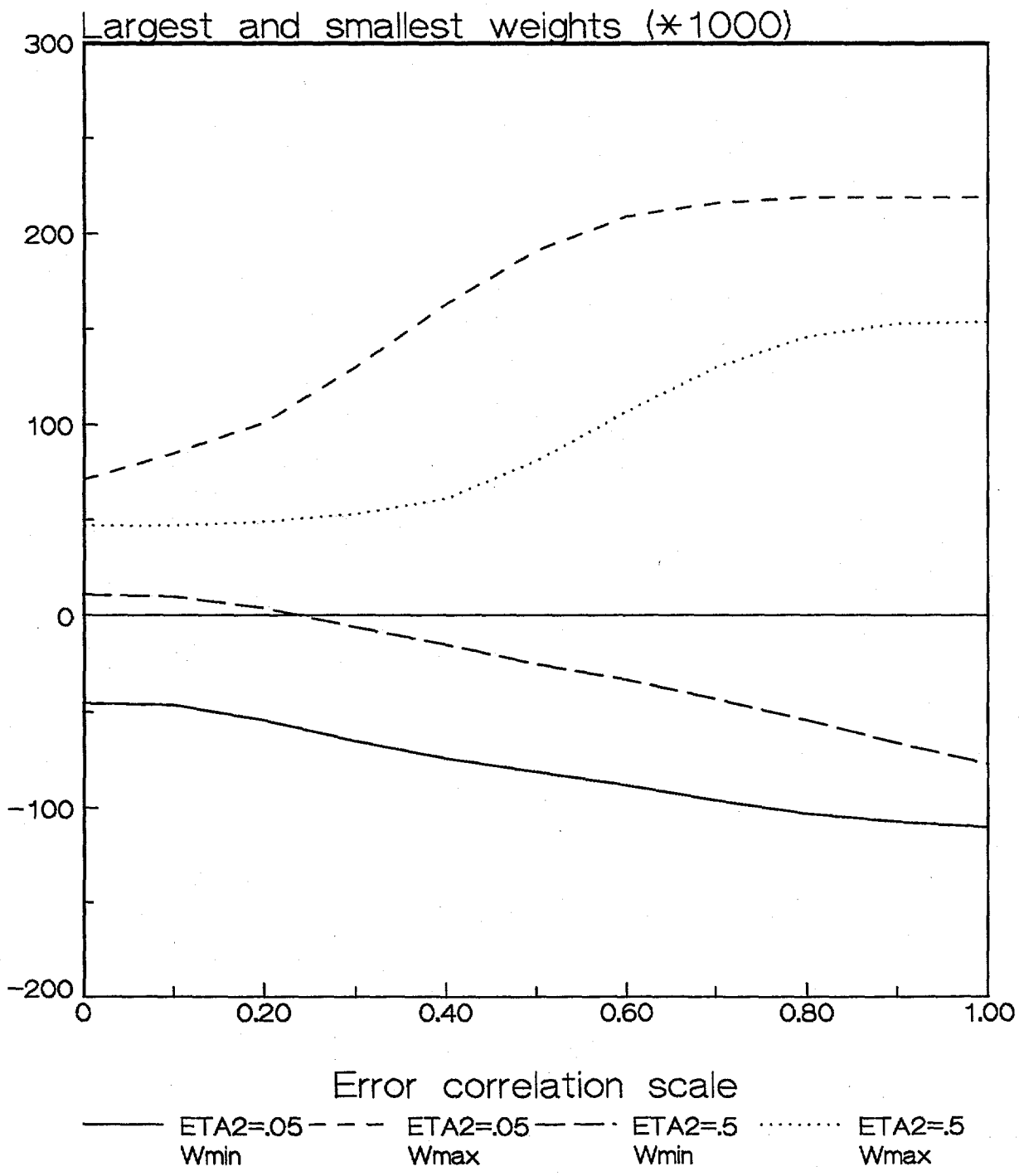


Fig. 23. Averaging of observations
with correlated errors



6. NON-OPTIMALITY OF AREA AVERAGING CAUSED BY APPROXIMATE KNOWLEDGE OF THE UNDERLYING STATISTICS.

Like any other statistically optimal procedure, the optimal averaging assures the smallest RMS error of the averaging *if the underlying statistics exactly reflect the reality*. The latter is never the case. Hypotheses of homogeneity and isotropy of correlation functions are approximate by their very nature, as is the hypothesis of stationarity of these functions in time. Moreover, even if we assume that these simplifying hypotheses are valid, it is necessary to take into account that our information on variances, correlation functions and random observational errors was obtained from some sample and is, like any statistical information, subjected to sampling errors. As a result, the statistics used in the averaging procedures may substantially differ from real ones, and the procedures may therefore not lead to the minimal error.

To stress this point, Norman Phillips proposed in his famous paper on the optimal interpolation (Phillips, 1976) to use the Russian superlative form *optimalneishy* for the estimates of the optimal interpolation accuracy under the assumption that the statistics used precisely describe the reality. He also analysed the consequences of the fact that the optimal interpolation is never optimalneishy.

Another wording, applied by some specialists, is to use the term statistical interpolation instead of optimal (or *optimum*, as proposed by Norbert Wiener, 1949) interpolation. Unfortunately, this terminology leads sometimes to a false impression among non-specialists that results of such an interpolation are expressed in statistical terms, while in fact, they are, of course, univalued.

The influence of inaccuracies in statistical information on the optimal interpolation results was investigated both theoretically and empirically comparatively long ago (e.g., Gandin, 1963). The overall conclusion is that this factor may cause a substantial decline in

the interpolation accuracy, particularly if the interpolation is multivariate. Recent investigations by A. Hollingsworth and P. Lonnberg (1986) and by H. Mitchell et al. (1990) confirmed this conclusion: they demonstrated that improvements in the approximation of statistical structure result in marked improvements in objective analysis and, thus, in numerical weather prediction.

Considering the same problem for the optimal averaging, it is necessary to realize, first of all, that all results presented in Sections 4 and 5 of this Office Note are valid under an idealized assumption that the information on statistical structure and random observation errors of the parameter in question, which is used in the course of the optimal averaging and the accuracy estimates, exactly reflects reality. In order to quantitatively evaluate the influence of this assumption, one can perform the following numerical experiments.

Let us assume that some statistics, used for the computation of optimal averaging weights, were different from the actual ones (which, just for these experiments, are also assumed to be known). Such averaging is thus not an optimal one, and its RMS error should be therefore computed by the "complete" equation (16), rather than by (20). To procede this way, we have first to construct a system (19), using the wrong μ, ζ and/or η. Solution of this system gives us the vector of *wrong weights* w_i. This weights should be then substituted into the equation (16) along with *correct statistics*, in order to compute the RMS error of such, non-optimal, averaging. Finally, we have to compare this error with that obtained from equation (20) (or 16) using *correct weights*, i.e., those obtained from the system (19) with correct statistics.

Results of some computations of this kind are shown on Figs 24 and 25. Only the RMS observation error η was assumed to be wrong in experiments resulted in Fig. 24, while the correlation function $\mu(r)$ was assumed to be correct. The two curves on Fig. 24 are for our two "standard" values of correct $\eta^2 = .05$ (modelling the use of climatological

norms as the background field) and .5 (for the use of forecast first guess). As the lower curve shows, an error in η does not result in any substantial loss in the averaging accuracy as long as the actual (normalized) RMS observation error is small. The only danger, demonstrated by the upper curve on Fig. 24, may arise when η is rather large and strongly underestimated when computing weights. This means, particularly, that if averaging deviations from the forecast first guess, one should assign a proper value of η, much larger than that for anomalies. It is safer under such circumstances to overestimate η than to underestimate it.

The second series of numerical experiments reported here dealt with erroneous values of the correlation radius, R (which leads to errors not only in μ, but in ζ and β as well) while the η value was assumed to be correct. Some results of these experiments are presented on Fig. 25. The independent variable in this figure is a correlation scale parameter SP defined as

$$SP = \begin{cases} 1 - \dfrac{R_c}{R_a} & R_a < R_c \\ \dfrac{R_a}{R_c} - 1 & R_a > R_c, \end{cases} \qquad (48)$$

where R_c=1 is the correct value of the correlation radius, and R_a is its assumed value for the weight computation. Positive SP values thus correspond to overestimated correlation radius, while negative ones reflect its underestimation.

It is seen on Fig. 25 that very large errors in the assumed correlation radius may substantially diminish the optimal averaging accuracy, particularly when the normalized RMS observation error is small. Errors of such size, however, practically never happen. So, points with SP= $\pm$3 on Fig. 25 correspond to the cases when the correlation radius is overestimated or understimated four times, while in practice, even a 50% error in the

correlation radius (reflected by |SP|=0.5) is considered as very large one. One may add to this, that errors in the correlation radius represent an extreme case of incorrect correlation statistics. What happens more often is an erroneous representation (e.g., approximation) of the correlation function with the same, or almost the same radius of correlation. The influence of such errors is, naturally, less than of those considered above.

The general conclusion from the estimates presented in this section is that the influence of inaccuracies in underlying statistics on the optimal averaging is much weaker than is the case for the optimal interpolation and, particularly, for the optimal differentiation. This conclusion is by no means unexpected: it is natural that the spatial integration diminishes sensitivity to statistical structure, while the spatial differentiation increases this sensitivity.

Fig. 24. Area-averaging accuracy with incorrect ETA2

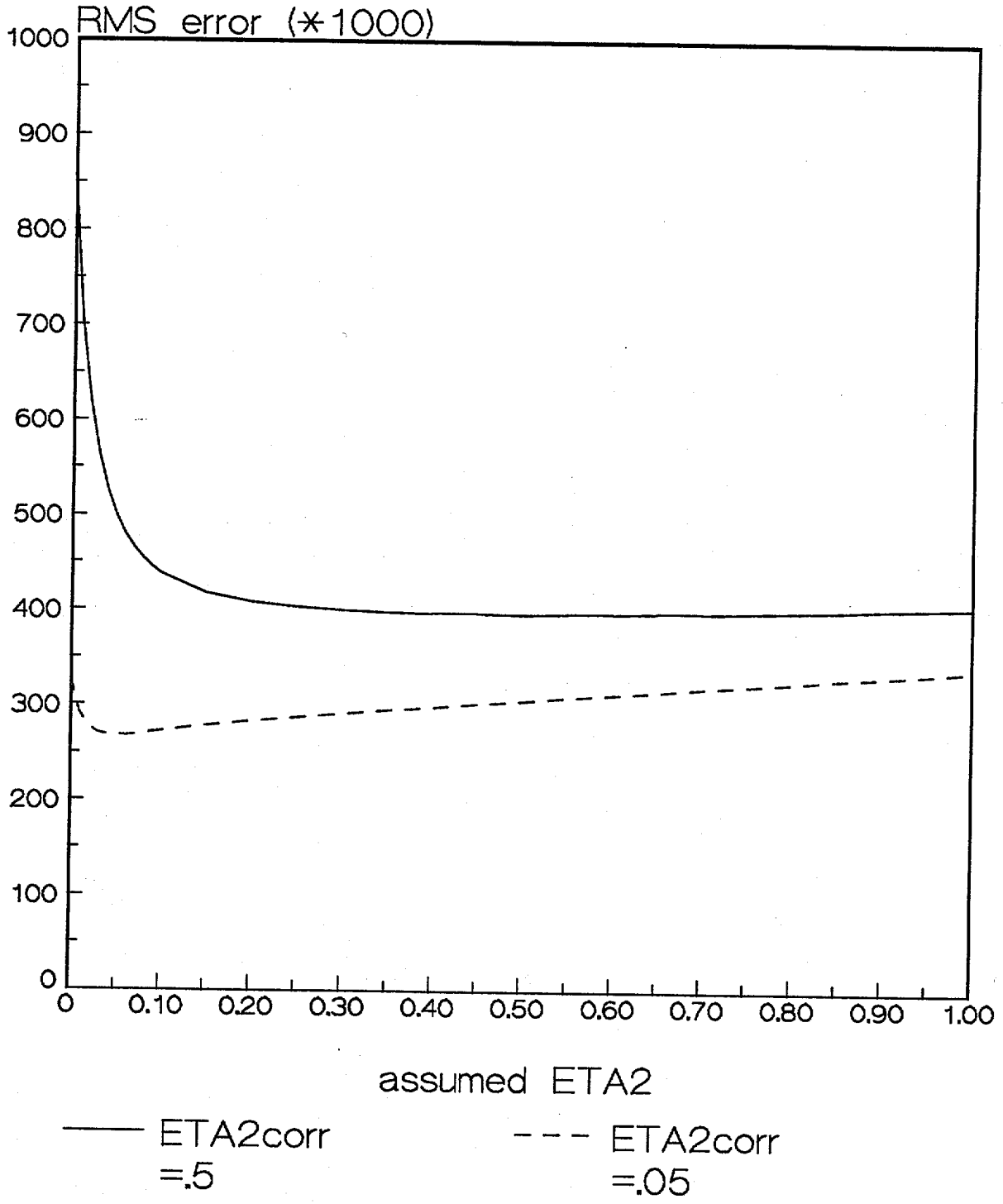
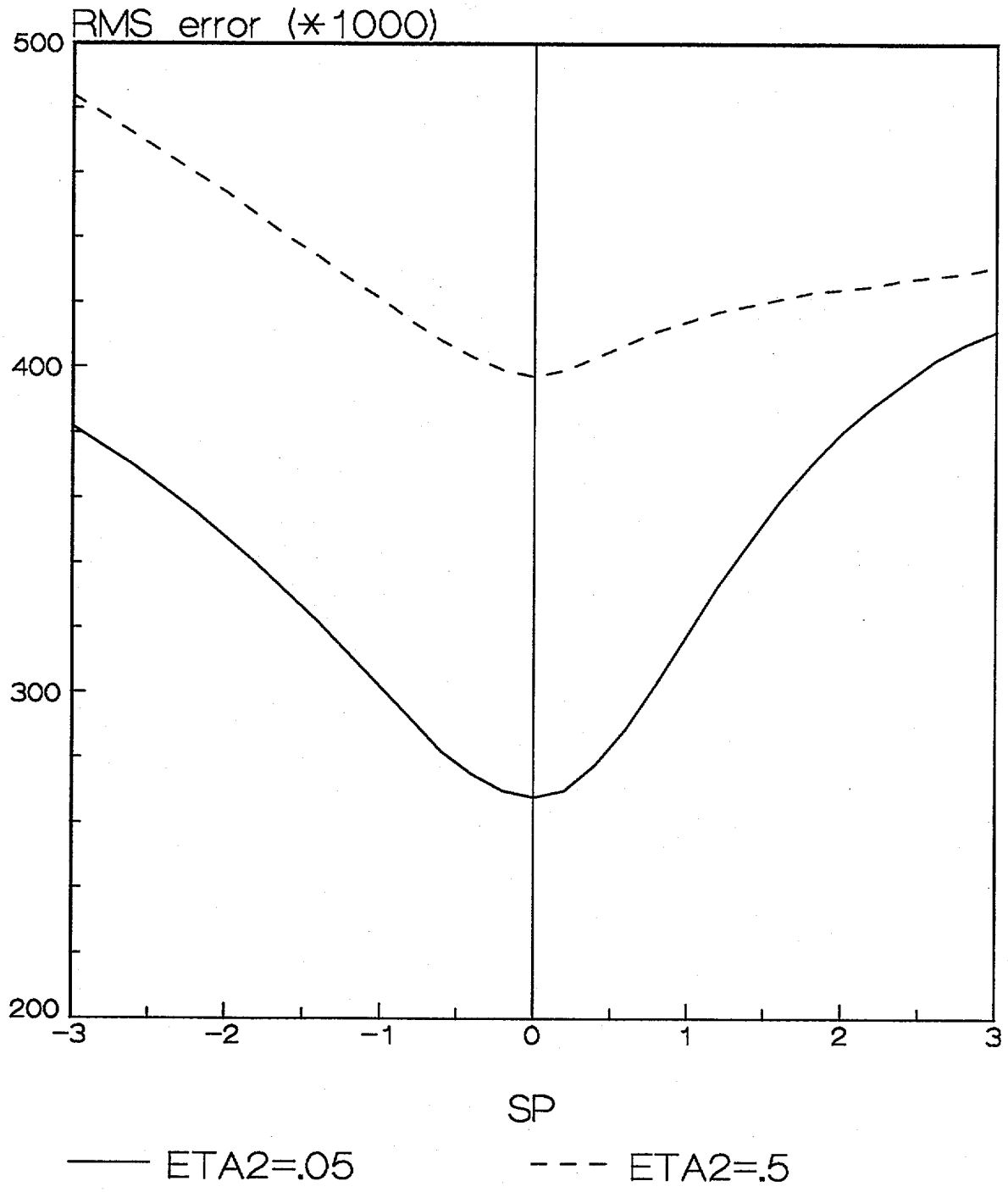


Fig. 25. Area-averaging accuracy with incorrect correlation radius



7. DISCUSSION AND CONCLUSIONS.

Although the area-averaging model used in this Office Note is rather simplified, there is little doubt that basic conclusions following from results of our computations with this model are applicable to much more general conditions of averaging a meteorological parameter over an area. Many of these conclusions are known from previous investigations and summarized in K79.

There is, however, at least one new topic, which has never been considered before. This is the application of forecast first guess as the background field in the course of area-averaging. All methods used so far for this purpose, including the optimal averaging, dealt either with anomalies or with "full" values. Moreover, initial values, subjected to area averaging, were usually those already averaged in time, like monthly mean values or anomalies.

No objections should be made, in principle, against spatial averaging of temporally averaged values, particularly if one is interested in detecting a small signal like the climate change. However, the statistical information, which is needed in order to estimate the accuracy of averaged values and/or to perform the optimal averaging is inevitably based on much smaller samples for temporal means than for instantaneous values and is therefore subjected to higher sampling errors. In addition to that, the instantaneous area-averaged values may be of interest for some problems other than that of detection the climate change.

Much more important, however, is the mere fact that the area averaging accuracy of increments (i.e., deviations from the forecast first guess) is substantially higher than that of anomalies, just like this is true for the optimal interpolation used in objective analysis of meteorological fields. It is necessary to stress in this respect that, while a simple arithmetic averaging of anomalies may lead to an acceptable estimate of the area-mean

value, particularly over a small domain, there is practically no alternative to the optimal averaging when the increments are being dealt with.

One certainly has to realize that the application of the forecast first guess as a background field makes the results of optimal averaging model-dependent, just as is the case with the optimal interpolation. Being practically negligible over regions with sufficiently dense observation network, this factor can play an important role if the density of observations or their accuracy is very low. This may be particularly dangerous for the problem of detecting any climate change (independently of whether the averaging is or is not applied for that): systematic errors in the numerical forecasts--the so-called biases--may create a false impression about climate changes over a data-poor region. As long as a bias does not persist over a large area, the averaging diminishes this danger, but it may still exist.

Another complication to be mentioned in this respect is that the RMS errors of averaging the increments are lower than those for full fields because of the random errors in the forecast first guess. Still, it is well known that the optimal interpolation of increments results in more accurate objective analyses of full fields than that of anomalies. There is little doubt that the same is true for the optimal averaging.

The area averaging of increments may be, at least in principle, performed in the course of operational objective analysis. It is much more convenient, however, to include it into the Climate Data Assimilation System (CDAS), now under design at NMC, because the CDAS is free of severe time limitations accompanying operational procedures. Even more promising, particularly for the problem of climate change, is the incorporation of optimal area-averaging procedures into the NMC Reanalysis (Kalnay and Jenne, 1991).

The idea of reanalysis is to perform anew objective analyses of past data applying a modern data assimilation system which uses an advanced numerical weather prediction model to produce first guess fields for all analyses. Undoubtedly, the application of the

Reanalysis System to a sufficiently long time sequence of global meteorological data is the best way to empirically investigate the global climate change, and the optimal averaging over some selected areas, or even over the whole globe, may be used in order to facilitate the detection of small climate change on the background of high natural weather variability. The availability of forecast first guess fields in the course of reanalysis allows us to perform the optimal averaging of increments and thus to achieve a higher accuracy.

One has to realize, when designing such procedures, that the optimal area-averaging is a more complicated process, at least from technical point of view, than the optimal (or any other) interpolation into points of a regular grid performed in the course of objective analysis. There is a natural way to limit the order of matrices for the optimal interpolation by selecting data from comparatively few observation points surrounding a grid point, and nothing like that exists for the optimal averaging. The order of matrices involved in the latter may be therefore quite high, particularly for averaging over large areas.

Two methods to avoid this difficulty have been proposed and applied. One approach is to perform first the interpolation into a regular grid and then to average arithmetically the grid point values over the area. Some authors even call this approach the optimal averaging (e.g., Weber, 1992). One may argue that this two-step approach must lead to less accurate results, particularly for areas non-homogeneously covered by observations. However, as demonstrated by Kagan (K79), the results of such approach are usually very close, at least for anomalies, to those of the "one-step" optimal averaging. The main objection against the two-step approach is therefore that there are practically no ways, using it, to estimate the averaging accuracy, while such estimates are produced automatically, as a by-product, in the course of optimal averaging.

Another way, also investigated by Kagan (K79), may be called an averaging "by parts": in order to obtain a value averaged over a large domain, one may perform the

optimal averaging over several parts of the domain and then average the results arithmetically. Once again, such procedure must be, in principle, less accurate but, as shown in K79, its results for anomalies are usually close to those of the optimal averaging over the whole domain; and again, the main objection against the averaging by parts is that it is very difficult, if not impossible, to estimate the accuracy of its final results.

The information on the averaging accuracy is of particular importance for the problem of detecting climate changes because it allows us to estimate our degree of confidence in any detected change. It is highly desirable therefore, when dealing with this problem, to apply the averaging over selected areas directly and in one step. As to the danger caused by the application of high-order matrices, it is minimal for averaging increments because of strong diagonal predominance in matrices involved. One may add to this that a simple way exists to diminish the matrix order if the density of observation points is quite high over some part of the averaging domain (or over the whole domain): one can diminish the density over such regions by simply ignoring data at some points or superobing them. As follows from results presented in Section 5 of this Office Note, such a procedure will practically not influence the estimates of area-averaged values or of their accuracy.

Another complication of the optimal averaging in comparison to the optimal interpolation is caused by the fact that multi-dimensional integrals of the point-value correlation function are involved: the cross-correlation function between point and area-mean values is expressed by the two-dimensional integral, and the variance of the mean value - even by the four-dimensional integral of the point-value correlation function. In contrast with the "toy example" considered in this Office Note, the only way to practically compute these integrals is to apply numerical quadrature methods. It is worthwhile to mention in this respect that, unless the averaging area is a circle, the cross-correlation function is non-isotropic, it depends on both coordinates of the point with respect to, say,

the center of the averaging domain. Numerical experiments, designed and performed recently by D. Deaven (pers. comm.), have demonstrated, however, that this does not create much problem if proper numerical methods are used. It will be also not difficult to include the area averaging of the forecast first guess fields, which is needed in order to obtain averaged values of parameter itself from those of its increments. One may use forecast data either in the form of spherical harmonic coefficients or at regular grid points to perform these computations.

A minor additional complication may be caused by the fact that the variance of the parameter in question is often not constant over a selected area, but varies geographically within the area. The so-called generalized homogeneity and isotropy hypothesis should be then used. It assumes that only the correlation function of such parameter is homogeneous and isotropic, while the covariance function is not. It is not clear to what extent this effect is important for performing the area averaging and/or for estimating its accuracy. In any case, however, it will be not difficult to correspondingly generalize the optimal averaging algorithm.

The numerical computations outlined may be essentially simplified if the averaging will be always done over the same set of domains, as it will be the case for CDAS and Reanalysis system. For each domain, the equation, expressing area-mean first-guess value in terms of harmonic coefficients or grid-point values, will be universal, not depending on time (as long as the model horizontal resolution remains unchanged). Even more important, the four-dimensional integral expressing the variance of area-averaged values needs to be computed only once for the whole period of stationary statistics (like a month or a season). The same may be done, at least in principle, for the cross-correlation (or normalized covariance) function: it may be computed beforehand and approximated in one or another way or tabulated for its use in the averaging. Most probably, however, this simplification will not be needed because, as D. Deaven demonstrated, the two-

dimensional quadratures involved in computations of the cross-correlation may be performed quite fast.

The most unclear is the situation with the temporal averaging of area-averaged values that should be performed in the course of reanalysis or after it. The optimal averaging in time is quite analogous, by its properties, to the area averaging and is technically simpler because of its one-dimensionality. What is needed, however, is the temporal averaging of area-averaged values (or vice versa). Results presented and discussed in this Office Note suggest that a simple arithmetic temporal averaging of values averaged over large areas can provide sufficient results, although a more detailed study is desirable, particularly for the averaging of increments. The point is, however, that we need not only such twice averaged values themselves, but also estimates of their accuracy.

This would not pose a problem if reliable information on the spatial-temporal statistical structure of meteorological parameters were available. Unfortunately, this is not the case. Almost all existing data of that kind are obtained using so-called separation of variables (or factorization). In other words, the spatial-temporal correlation is assumed to be a product of spatial and temporal correlations. Independently of the degree of accuracy achievable with such approximation, it hardly can be used for our purposes: it immediately follows from the separation assumption that the temporal correlation between spatially averaged values does not differ at all from the temporal correlation of point values (and vice-versa). At the same time, it is intuitively clear that the spatial averaging should increase the temporal correlation, although the quantitative aspect of this effect is not well known.

Perhaps, the best way to approach this problem is just to investigate the statistical structure of area-mean values in the course of reanalysis (and/or when applying the CDAS). Information obtained this way may be then used in order to estimate the

accuracy of spatial-temporal averaging and, maybe, even to improve the temporal averaging of spatially averaged values.

The last, but far not the least important, question to be discussed here is which parameter or parameters to average. The methodology outlined in this Office Note is applicable to any scalar parameter whose statistical structure may be assumed homogeneous and isotropic over an area in question and is known, as is the case for temperature and isobaric height. It will be not difficult to develop a slightly more general method allowing the area averaging of vector fields, like the wind field.

The most challenging task is, however, to average those meteorological and hydrological parameters for which the point values by themselves are not representative enough, like humidity and, particularly, precipitation. This is far from being a simple problem, not only because statistical structure of such parameters is not well known, but also because their probability distribution essentially differs from the normal (Gaussian) one. Strictly speaking, the normal distribution is a necessary condition for any method based on the mean square error minimization. At the same time, the probability density curve for precipitation sums over a small interval, like 12 or 24 hours, is not only far from normal, it usually has a singularity at zero. This means that the precipitation values should be first summed up (or averaged) over larger intervals of time, and their statistics should be investigated and applied only after that. Some investigations of this kind have been already undertaken (see, e.g., K79).

Another interesting possibility is to apply the area averaging method not to observed quantities themselves, but to some of their vertically integrated values, like the so-called precipitable water, which is a vertical integral of humidity, or layer-integrated horizontal mass fluxes. The area averaging may be also applied to various differential characteristics, like vorticity and divergence. Statistics needed for this may be derived from those for the wind vector. One may mention in this respect that the area-mean

vorticity is proportional to velocity circulation along the boundary of the area and, analogously, the area-mean divergence is proportional to the mass flux across the boundary. These quantities can be also integrated along the vertical.

The general conclusion from this discussion is that there exist many possible applications of the optimal averaging in the course of CDAS and Reanalysis projects. The task is to select most promising among such applications, particularly for the problem of global climate change.

Aknowledgments.

I would like to express, first of all, my gratitude to Eugenia Kalnay for her enthusiastic and constructive support of investigations resulted in this Office Note. Many colleagues at NMC and elsewhere, particularly Dennis Deaven at the NMC Development Division, Leonid Rukhovets, Dick Reynolds ant Tom Smith at CAC, Rol Madden at NCAR and Marvin Geller at SUNY, demonstrated their interest in the optimal averaging method and applied it in their work. I am particularly grateful to Bill Collins for his permanent support and for his important help in the preperation of this Office Note.

REFERENCES.

Budyko, M.I., 1982. The Earth's climate: past and future.

Academic Press, 307 pp.

Daley, R., 1991. Atmospheric data analysis. Cambridge University Press. 457 pp.

Gandin, L.S., 1965. Objective analysis of meteorological fields. Translated from Russian. Israel Program for Scientific Translations, Jerusalem. 242 pp.

Gandin, L.S., 1970. The planning of meteorological station network. WMO Technical Note No 111, 42 pp.

Hollingsworth, A. and P.Lonnberg, 1986. The statistical structure of short-range forecast errors as determined from radiosonde data. Tellus, vol. 38, 111-136 (Part 1), 137-161 (Part 2).

Kagan, R.L., 1979. Averaging of meteorological fields. Gidrometeoizdat, Leningrad. 214 pp. (in Russian).

Kalnay, E. and R.Jenne, 1991. Summary of the NMC/NCAR reanalysis workshop of April 1991. BAMS, vol. 72, No 12, 1897-1904.

Mitchell, Herschel L. et al, 1990. Revised interpolation statistics for the Canadian data assimilation procedure: their derivation and application. Mon. Wea. Rev., vol.118, No 8, 1591-1604.

Phillips, N.A., 1976. The impact of synoptic observing and analysis systems on flow pattern forecasts. BAMS, vol. 57, No 10, 1225-1240.

Thiebaux, H.J., M.A.Pedder, 1987. Spatial objective analysis with applications in atmospheric science. Academic Press, 299 pp.

Vinnikov, K., P.Ya.Groisman and K.M.Lugina, 1990. Empirical data on contemporary global climate changes (temperature and precipitation). Journ. Climate, vol. 3, 662-677.

Weber, R., 1992. Statistically optimal averaging for the determination of global mean temperature. 5th International Meeting on Statistical Climatology, Toronto, Canada, 421-424.

Wiener, N., 1950. Extrapolation, interpolation and smoothing of stationary time series. New York, London. 163 pp.