SAND2000-1708J

A New Classical Least Squares/Partial Least Squares Hybrid Algorithm for Spectral Analyses*

David M. Haaland and David K. Melgaard

Sandia National Laboratories

Albuquerque, New Mexico 87185-0342

ABSTRACT

A new classical least squares/partial least squares (CLS/PLS) hybrid algorithm has been developed that demonstrates the best features of both the CLS and PLS algorithms during the analysis of spectroscopic data. By adding our recently reported prediction-augmented (PACLS) features to the hybrid algorithm, we have the added ability to incorporate known or empirically derived spectral shape information into the hybrid algorithm to correct the hybrid model for the presence of unmodeled sources of spectral variation. The detailed steps of the new hybrid algorithm in calibration and prediction are presented. The powerful capabilities of the new PACLS/PLS hybrid are demonstrated for the near-infrared spectra of a system of dilute aqueous solutions containing the analytes urea, creatinine, and NaCl. The PACLS/PLS method is used to correct the detrimental effects of unmodeled solution temperature changes and spectrometer drift in the multivariate spectral calibration model. PLS and PACLS/PLS predictions of analytes from the variable-temperature data were made with models based upon spectra previously taken of the samples at constant temperature. Prediction errors were inflated by more than an order of magnitude due to the presence of unmodeled temperature variations and system drift. PLS achieved improved predictions of the variable-temperature spectra by adding spectra of a few variable-temperature samples

DISCLAIMER

into the original calibration data followed by recalibration. PACLS/PLS predictions were corrected for temperature variations and system drift by adding spectral differences of the same subset of samples collected under constant- and variable-temperature conditions to the PACLS prediction portion of the hybrid algorithm during either calibration or prediction. Comparisons of the prediction ability of the hybrid algorithm relative to the PLS method using the same calibration and subset information demonstrated hybrid prediction improvements that were significant at least at the 0.01(proper terminology?) level for all three analytes. The new hybrid algorithm has widespread uses, some of which are also discussed in the paper.

*Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-ACO4-94AL85000.

INTRODUCTION

Multivariate spectral calibrations are now standard methods for performing quantitative spectral analyses. Partial least squares (PLS) and principal component regression (PCR) are the most common multivariate methods for quantitative spectral analyses.[1,2,3,4,5,6] Although classical least squares (CLS) and inverse least squares (ILS) methods were used extensively before 1990,[7,8,9] they are now used less often. However, Haaland and coworkers[7,10] have advocated the use of CLS multivariate calibration for qualitative interpretation of spectral data. CLS also has advocates in the infrared analysis of gas-phase samples[11,12,13] and in the analysis of atomic emission spectra.[14,15,16] Martens and Naes[17,18] have reported an extended Beer's law model that greatly improves the applicability of CLS, but this method has not received much attention in the literature. Recently, Haaland and Melgaard have presented an enhanced version of CLS, named prediction-augmented CLS (PACLS),[19] that further improves the extended Beer's law model of Martens and Naes and demonstrates significant advantages relative to standard CLS methods and is less restrictive than CLS. With the PACLS algorithm, known spectral information about interfering spectral components can be efficiently included in the PACLS algorithm to readily correct for their presence in the calibration or unknown sample spectra. The restriction on the CLS algorithm that the concentrations of all spectrally active components must be included during calibration is relaxed with PACLS if spectral information of interfering spectral components is added during CLS prediction. In the remainder of this document, we will refer to the added spectral information as "spectral shapes" since the example here uses a single continuous spectral region in the analyses. However, the analyses can also be performed using multiple spectral regions or

even multiple isolated spectral intensities where the term "spectral shapes" has less meaning.

We have also demonstrated that the PACLS algorithm has the ability to accommodate the presence of unmodeled components in unknown samples if the spectral shape of the unmodeled component can be determined and added during the CLS prediction step.[19] Additional uses of the PACLS algorithm include correction of the effects of spectrometer drift and changes between spectrometers with the use of repeat sample spectra and subset sample spectra, respectively. Thus, PACLS has the ability to address maintenance and transfer of calibration. However, the PACLS algorithm is still too restrictive in many real cases encountered in industry since it requires that all spectral components be included as concentrations during calibration or added as spectral shapes during prediction. Historically, PLS is has gained widespread use because it has the flexibility to define a calibration model even when all the spectrally active components are not known.

In an effort to take advantage of the qualitative interpretation of CLS and the special capabilities of PACLS outlined above while retaining the flexibility of PLS, we have developed a new hybrid algorithm that has the advantages of CLS, PACLS, and PLS. The hybrid method is named CLS/PLS hybrid or PACLS/PLS hybrid depending on whether CLS or PACLS is used during the initial steps of the hybrid algorithm. We hypothesize that the new PACLS/PLS algorithm will have advantages over existing methods in cases involving maintenance and/or transfer of calibration or when unmodeled species are present in the unknown samples to be predicted. In this paper, we apply the new hybrid algorithm to the quantitative analysis of a system of dilute aqueous

solutions. The power of the method is demonstrated for the case where a constant-temperature model is applied to variable-temperature data. We demonstrate that the new hybrid method can handle the presence of the unmodeled temperature variation and spectrometer drift with the use of a set of subset sample spectra obtained under both constant- and variable-temperature conditions. Alternatively, we can independently determine the spectral effects of temperature changes and system drift on the sample spectra and then add the appropriate spectral shapes to the PACLS/PLS hybrid algorithm. The new algorithm yields better prediction ability than CLS, PACLS, or PLS applied to the same data. It also outperforms the procedure of simply adding the data from the variable-temperature subset samples back into the constant-temperature data followed by recalibration using the standard PLS algorithm. Our hybrid method yields improved prediction precision relative to the newly described synthetic PLS method by more effectively accommodating for the presence of unmodeled spectral interferences.[20] The new hybrid algorithm is efficient and can eliminate the need for collecting a new calibration set when the unknown samples exhibit unmodeled sources of spectral variation.

EXPERIMENTAL METHODS

The samples used in this study are a series of 31 dilute aqueous solutions of urea, creatinine, and NaCl. These samples and the experimental details have been presented previously.[19,20,21] The samples were prepared by weight and volume in a pseudo D-optimal design[22] that allowed each of the three analytes to be varied separately at 16 levels from 0 to approximately 3000 mg/dL. The sample solutions were sealed along

with a magnetic stirring bar in 10-mm pathlength cuvettes. All sample spectra plus single repeat spectra of 3 samples were collected on one day at a constant temperature of 23°C. Several days later, the spectra of each sample were collected under variable temperature conditions from 20 to 25°C in 1°C intervals. The temperatures were held constant to 0.05°C ($\pm 1\ \sigma$) with a HP temperature controller that included a magnetic stirrer. Spectra of pure water were also obtained in random order at 0.5°C intervals from 20 to 25°C in an experiment run on a separate day.

The near-infrared spectra were obtained on a Nicolet Model 800 Fourier transform infrared (FT-IR) spectrometer. The spectrometer employed a 75 W tungsten-halogen lamp, quartz beam splitter, and liquid-N_2-cooled InSb detector. Spectra at 16 cm^{-1} resolution were obtained with 256 scans signal averaged. Although background spectra of the empty cuvette were collected after each sample, best prediction results were obtained by using an average background obtained from all the background single-beam spectra collected during the day. Analyses were always performed after converting the ratioed spectra to absorbance. Spectra were analyzed over the spectral region from 7500 to 11,000 cm^{-1}.

The data were analyzed using PLS and PACLS/PLS hybrid algorithms incorporated into software developed at Sandia National Laboratories. The software was programmed using the Array Basic programming language from the GRAMS 32 software obtained from Galactic Industries Corporation. Cross-validated calibrations were performed removing all the data from a given sample during each rotation. Since the calibration data contained repeat spectra from three samples, 31 sample rotations were performed in these analyses. Previously reported calibrations[19-21] did not include

the 3 spectra from the constant-temperature repeat samples. Two sets of five subset samples were selected from the variable-temperature data for augmenting the constant-temperature PLS and hybrid calibration data. The first set of 5 samples had temperature variations and maximally spanned the concentration variation of calibration samples. The second set included 5 of the 6 samples that were obtained at 23°C during the variable-temperature data collection, i.e., samples that were collected at the same temperature as the constant-temperature sample set. These subset sample data were used to separately augment the constant-temperature PLS and PACLS/PLS models.

THEORY

The first two steps of the PLS algorithm have been explained by Haaland and Thomas[2] as CLS calibration and CLS prediction for one component. The new hybrid algorithm adds a completely separate CLS calibration and prediction front end to the full PLS model. Thus, the new hybrid method might be considered an extension of the first two single-component CLS steps by allowing the initial CLS calibration step of PLS to use all concentrations of species that are known in the calibration samples. This first improvement has the advantage of generating pure-component spectral estimates that are closer to the true pure-component spectra than are obtained from the first PLS weight-loading vector since the influence of any known interferences are minimized with the more complete CLS calibration step of the hybrid algorithm. The addition of more component information in the first CLS step should also generate a more parsimonious model than PLS since the first weight-loading vector of the hybrid is rotated in a direction that is closer to the pure-component spectrum of the analyte. A similar analogy

to the parsimony improvement of PLS over PCR has been made previously[2,3,17] based on the argument that the first weight-loading vector of PLS is more similar to the pure-component spectrum of the analyte relative to the first eigenvector of the PCR model.

Also, the additional powerful capability of prediction-augmentation used in PACLS can be added to the CLS prediction step of the new hybrid algorithm. That is, any analytes or spectral effects whose concentrations or reference values, respectively, are not known in the calibration can have their pure-component spectral shapes added during the CLS prediction step of the hybrid modeling. Here the "spectral shapes" added in the CLS prediction step do not have to be representative of molecular components. They can in fact represent the spectral effects of spectrometer drift, the spectral changes caused by changing spectrometers, or the effect of other influences on the spectra such as sample temperature, purge gas variations, or optical effects due to sample insertion variations. Thus, the appropriate augmentation in the second CLS prediction step of the hybrid algorithm can provide for maintenance of calibration, transfer of calibration, or accommodating the influence of spectral variations or spectral components not included in the calibration spectra. The concentration and spectral residuals from the CLS calibration and CLS or PACLS prediction are then passed to the PLS portion of the hybrid algorithm as the starting concentrations and spectra. With the PLS portion of the hybrid algorithm, we build a model based on CLS residuals that can correct the CLS concentration estimates for any sources of spectral variation left out of the CLS portion of the model.

In the description of the hybrid calibration and prediction algorithms, matrices are represented by bold upper-case letters, vectors as lower-case bold letters, and scalars are

in italics. Vectors are represented as column vectors and row vectors are described as transposed column vectors. The superscript T indicates transposed matrices and vectors, the superscript –1 is used to indicate the inverse, and the superscript + represents the pseudoinverse matrix.[23] The pseudoinverse can be solved by a variety of mathematical procedures such as QR decomposition or singular value decomposition.[23] In this paper, the generic hybrid algorithm will be called CLS/PLS when there is no prediction augmentation step. We further distinguish the hybrid by adding a 2 to the acronym to indicate when the concentrations of 2 or more components are included in the CLS calibration portion of the algorithm (similar to the PLS2 notation for PLS when 2 or more analytes are simultaneously included in the analysis). CLS1/PLS will be used to indicate the hybrid algorithm where there is no prediction augmentation and the analysis is preformed using concentrations from a single component. PACLS2/PLS and PACLS1/PLS are the comparable hybrid algorithms where the prediction step of the CLS portion of the hybrid is augmented with spectral shape information. Note that the prediction augmentation can occur during either the hybrid calibration or the hybrid prediction or during both calibration and prediction phases of the algorithm. PACLS/PLS refers generically to either PACLS1/PLS or PACLS2/PLS. The term "hybrid algorithm" will encompass the full family of hybrid models listed above.

Hybrid Calibration Algorithm

Step 1. Pretreatment of data.

Center **A** and **C** (optional) where **A** is an $n \times p$ matrix of the n sample spectra with p spectral intensities. **C** is an $n \times m$ matrix of the reference concentrations for the n

samples and m components. If all spectrally active chemical components are included in **C** and each row of **C** adds to 1, a singularity in the least-squares solution will be generated. To avoid such a singularity in step 2, either leave out one component in **C** or do not mean center the data.

Scale **A** (optional).

Set index h to 1.

Step 2. Estimate the CLS pure-component spectra ($\hat{\mathbf{K}}$) where **K** is an $m \times p$ matrix and $\mathbf{E_A}$ is an $n \times p$ matrix of spectral errors.

Model: $$\mathbf{A} = \mathbf{CK} + \mathbf{E_A} \quad (1)$$

Least-squares solution: $$\hat{\mathbf{K}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{A} = \mathbf{C}^+\mathbf{A} \quad (2)$$

Step 3. Estimate the CLS concentrations using PACLS prediction.

Add spectral baseline components and additional known spectral features as rows to $\hat{\mathbf{K}}$.

Increase the size of the **C** and $\hat{\mathbf{C}}$ matrixes to accommodate the estimated concentrations for the original plus the augmented components. Label these augmented matrixes $\hat{\tilde{\mathbf{K}}}$, $\tilde{\mathbf{C}}$, and $\hat{\tilde{\mathbf{C}}}$, respectively.

Model: $$\mathbf{A} = \tilde{\mathbf{C}}\,\hat{\tilde{\mathbf{K}}} + \mathbf{E_A} \quad (3)$$

Least-squares solution: $$\hat{\tilde{\mathbf{C}}} = \mathbf{A}\hat{\tilde{\mathbf{K}}}^T(\hat{\tilde{\mathbf{K}}}\hat{\tilde{\mathbf{K}}}^T)^{-1} = \mathbf{A}(\hat{\tilde{\mathbf{K}}}^T)^+ \quad (4)$$

Step 4. Calculate CLS residuals.

Spectral residuals: $$\mathbf{E_A} = \mathbf{A} - \hat{\tilde{\mathbf{C}}}\hat{\tilde{\mathbf{K}}} \quad (5)$$

Concentration residuals: $$\mathbf{E_C} = \hat{\mathbf{C}} - \mathbf{C} \quad (6)$$

where $\hat{\mathbf{C}}$ is derived from $\hat{\tilde{\mathbf{C}}}$ by removing the columns of the $\hat{\tilde{\mathbf{C}}}$ matrix corresponding to the added spectral shapes.

Step 5. Isolate the results for a single analyte from $\mathbf{E_C}$ to form the concentration residuals $\mathbf{e_c}$ for that analyte.

The following PLS-like steps are to be performed one selected analyte at a time.

Set $h = 1$

Step 6. Estimate weight-loading vector ($\hat{\mathbf{w}}_h$) for selected analyte.

Model: $\mathbf{E_A} = \mathbf{e_c}\mathbf{w}^{\mathrm{T}} + \mathbf{E'_A}$ where $\mathbf{E'_A}$ is used to differentiate the PLS spectral residuals from the CLS residuals $\mathbf{E_A}$.

Least-squares solution:
$$\hat{\mathbf{w}}_h = \frac{\mathbf{E_A}^{\mathrm{T}}\mathbf{e_c}}{\mathbf{e_c}^{\mathrm{T}}\mathbf{e_c}} \tag{7}$$

Normalize $\hat{\mathbf{w}}_h$

Step 7. Form score vector ($\hat{\mathbf{t}}_h$) for selected analyte.

Model:
$$\mathbf{E_A} = \mathbf{t}_h\hat{\mathbf{w}}_h^{\mathrm{T}} + \mathbf{E'_A} \tag{8}$$

Least-squares solution:
$$\hat{\mathbf{t}}_h = \frac{\mathbf{E_A}\hat{\mathbf{w}}_h}{\hat{\mathbf{w}}_h^{\mathrm{T}}\hat{\mathbf{w}}_h} = \mathbf{E_A}\hat{\mathbf{w}}_h \tag{9}$$

Step 8. Relate score vector ($\hat{\mathbf{t}}_h$) to the concentration residual.

Model:
$$\mathbf{e_c} = v_h\hat{\mathbf{t}}_h + \mathbf{e'_c} \tag{10}$$

where $\mathbf{e'_c}$ represents the new concentration residual for the PLS portion of the hybrid algorithm.

Least-squares solution:
$$\hat{v}_h = \frac{\hat{\mathbf{t}}_h^{\mathrm{T}}\mathbf{e_c}}{\hat{\mathbf{t}}_h^{\mathrm{T}}\hat{\mathbf{t}}_h} \tag{11}$$

Step 9. Formation of $\hat{\mathbf{b}}_h$, the PLS loading vector for the CLS spectral residuals, $\mathbf{E}_\mathbf{A}$.

Model: $$\mathbf{E}_\mathbf{A} = \hat{\mathbf{t}}_h \mathbf{b}_h^\mathrm{T} + \mathbf{E}'_\mathbf{A} \tag{12}$$

Least-squares solution: $$\hat{\mathbf{b}}_h = \frac{\mathbf{E}_\mathbf{A}^\mathrm{T} \hat{\mathbf{t}}_h}{\hat{\mathbf{t}}_h^\mathrm{T} \hat{\mathbf{t}}_h} \tag{13}$$

Step 10. Calculation of the PLS spectral and concentration residuals

Spectral residuals: $$\mathbf{E}'_\mathbf{A} = \mathbf{E}_\mathbf{A} - \hat{\mathbf{t}}_h \hat{\mathbf{b}}_h^\mathrm{T} \tag{14}$$

Concentration residuals: $$\mathbf{e}'_\mathbf{c} = \mathbf{e}_\mathbf{c} - v_h \hat{\mathbf{t}}_h \tag{15}$$

Step 11. Increment h, substitute $\mathbf{e}'_\mathbf{c}$ for $\mathbf{e}_\mathbf{c}$ and $\mathbf{E}'_\mathbf{A}$ for $\mathbf{E}_\mathbf{A}$ and repeat Steps 6 through 10 for the desired number of loading vectors.

Step 12. Repeat from Steps 5 to 11 for other analytes in the calibration samples.

Hybrid Prediction Algorithm

Step 1. Center the unknown sample spectrum, **a**, if **A** was centered in the calibration. Also scale **a** if **A** was scaled.

Step 2. If desired, augment the rows of $\hat{\tilde{\mathbf{K}}}$ further if new spectral shapes are to be added during prediction (e.g., to accommodate system drift with repeat sample spectral differences). Further augment the size of $\tilde{\mathbf{c}}$ to accommodate the predicted values that correspond to the spectral shapes added to $\hat{\tilde{\mathbf{K}}}$.

Step 3. Estimate the initial CLS-estimated concentrations for the analyte in the unknown sample.

Model: $$\mathbf{a} = \hat{\tilde{\mathbf{K}}}^\mathrm{T} \tilde{\mathbf{c}} + \mathbf{e}_\mathbf{a} \tag{16}$$

Least-squares solution: $$\hat{\tilde{\mathbf{c}}} = (\hat{\tilde{\mathbf{K}}} \hat{\tilde{\mathbf{K}}}^\mathrm{T})^{-1} \hat{\tilde{\mathbf{K}}} \mathbf{a} \tag{17}$$

Step 4. Calculate CLS spectral residuals using the model in Step 3.

$$\mathbf{e}_a = \mathbf{a} - \hat{\tilde{\mathbf{K}}}^T \hat{\tilde{\mathbf{c}}} \tag{18}$$

Step 5. Isolate analyte concentration $\hat{c}$ from $\hat{\tilde{\mathbf{c}}}$, add the average analyte concentration $\bar{c}$ from the calibration if the data were mean centered, and label the result $\hat{c}_0$, i.e.,

$$\hat{c}_0 = \hat{c} + \bar{c} \tag{19}$$

Set $h = 1$.

Step 6. Calculate the PLS score from CLS spectral residual

Solution: $$t_h = \hat{\mathbf{w}}_h^T \mathbf{e}_a \tag{20}$$

Step 7. $$\hat{c}_h = \hat{c}_{h-1} + \hat{v}_h t_h \tag{21}$$

Step 8. $$\mathbf{e}_h = \mathbf{e}_{h-1} - \hat{\mathbf{b}}_h t_h \tag{22}$$

Step 9. Increment h, substitute $\mathbf{e}_h$ for $\mathbf{e}_a$ and repeat Steps 6 through 9 until $h = r$, where r is the number of PLS factors in the hybrid model.

Note: $\hat{\mathbf{w}}_h$, $\hat{v}_h$, and $\hat{\mathbf{b}}_h$ are from the PLS portion of the hybrid calibration algorithm, and $\mathbf{e}_0 = \mathbf{e}_a$.

The above hybrid calibration and prediction algorithms represent the full calibration and prediction algorithms. However, a cross validation procedure that rotates out one or more sample spectra at a time is used for factor selection and improved outlier detection. The optimal number of PLS factors to use in the hybrid algorithm is determined using cross validation in the same manner as we proposed earlier for the PLS model factor selection.[2] It is important to note that the cross-validation procedure must be performed properly to obtain optimal factor selection. A full CLS calibration should not be executed first followed by cross-validated PLS applied to the residual concentrations and residual spectra from the full CLS model. In this latter case, PLS will

overfit the CLS residuals, and the predicted error sum of squared errors (PRESS) will continuously decrease with increased numbers of PLS factors. The entire hybrid algorithm must be part of the cross-validation procedure in order for factor selection to be optimized. The same outlier detection metrics such as spectral F ratio, concentration F ratio, Mahalanobis distance, and cross-validated spectral residuals described previously[2,24] can also be developed during the generation of the cross-validated hybrid model.

We can see from the hybrid calibration and prediction algorithms that there is an opportunity to add spectral shapes in either the calibration (during the CLS prediction step of the hybrid calibration) or the prediction portions of the algorithm or in both parts of the algorithm. In the case of the PACLS algorithm, the same sample prediction results are obtained independent of whether the spectral shapes are added during cross-validated calibration or during prediction. The hybrid algorithm, on the other hand, has a CLS prediction step in the calibration as well as during the prediction of unknown samples. Whether we add the spectral shapes during calibration or prediction affects the hybrid prediction results of the unknown samples. We find that best prediction ability is sometimes achieved by adding the spectral shapes during calibration. In addition, outlier detection is more effective when all the spectral shape information is included during calibration. However, updating the hybrid model during prediction is much faster computationally than updating the cross-validated calibration model.

As described in Ref. 19, the spectral shapes added in the PACLS part of the hybrid algorithm can represent pure spectral variations due to chemical components or any other sources of spectral variation (e.g. temperature changes, spectrometer drift,

purge gas variations, changes in spectrometers, etc.), or they can represent different linear combinations of the underlying sources of spectral variation. If linear combinations of pure spectral effects are used in the PACLS portion of the algorithm, we must add as many independent linear combinations as there are pure spectral effects present in the data in order to correct the model for all sources of spectral variation in the data. However, if quantitation of the pure spectral interference is not required, then the magnitude of its spectral shape is not important when adding spectral shapes to the PACLS portion of the hybrid algorithm.

RESULTS AND DISCUSSION

The individual constant- and variable-temperature spectral data sets, separately mean-centered, are shown in Fig. 1 along with the mean spectrum of all spectra. Much of the variance in the spectral data is a result of spectrometer drift. The drift present in the variable-temperature spectra is somewhat greater than in the constant temperature data either due to the longer data collection time for the variable-temperature experiments or due to greater system drift on the day of the collection of the variable-temperature data. The effects of temperature variation in the variable-temperature set are also evident as greater variance in the spectra on the high-energy side of the water absorption features.

Cross-validated prediction results for PLS, CLS/PLS with all known components and time of data collection added in the CLS portion of the calibration (CLS2/PLS), and CLS/PLS with only a single analyte added during the calibration (CLS1/PLS) are presented in Table I. Factor selection is based upon the method suggested by Haaland and Thomas.[2] As can be seen in Table I, all three methods yield comparable cross-

validated calibration prediction precisions. Thus, the new hybrid models do not exhibit significant cross-validated calibration improvements with this data set. Their advantage will be shown to be evident during true prediction. Figure 2 presents the plot of cross-validated calibration predictions versus reference values for creatinine using a CLS2/PLS model applied to the constant-temperature data.

If any of the above constant-temperature models are applied to the variable-temperature sample spectra, large prediction errors are encountered as demonstrated in Table II for all three models and presented for creatinine in Fig. 3 using the CLS2/PLS model. For comparison with later results using models that incorporate subset spectral information, the results in Table II and Fig. 3 do not include predictions of the 5 sample spectra that are used in later calibrations that employ the subset samples to adjust the models for temperature variations and system drift. Thus, the spectral effects of temperature variations and long-term system drift on the spectra are extremely detrimental to the predictions. The higher SEP's indicate that the hybrid model predictions may be even more sensitive to spectra outside the original calibration model than the PLS model predictions.

We have previously suggested methods of generating synthetic PLS calibration models[20] by adding variable amounts of the spectral shape of a temperature change in the solutions to the original constant-temperature spectra to allow temperature variations to be incorporated into the model. Another method to incorporate temperature variations into the constant-temperature model is to collect a subset of sample spectra that includes temperature variations and add these subset spectra to the original constant-temperature data. PLS models generated for the subset-augmented data set then contain information

about the interfering temperature variations and system drift. This augmented PLS method reduces the time and effort involved relative to regenerating the PLS calibration model using an entire new set of sample spectra collected under variable-temperature conditions.

We can compare predictions of variable-temperature spectra using the above recalibration method with the predictions obtained from PACLS2/PLS and PACLS1/PLS models where spectral shapes derived from the same subset samples are added during calibration. The spectral shapes added to the hybrid models are the pair-wise difference spectra from the five selected samples run both with the original constant-temperature data and the variable-temperature prediction samples. These difference spectra should not contain any analyte information since the composition of the samples is constant in the sealed cells. The difference spectra, therefore, contain linear combinations of the spectral variations due to within and between-day spectrometer drift and the temperature changes in the samples.

Comparisons are given in Table III for the prediction ability of the augmented PLS model and two hybrid models using the same calibration and subset spectral information in the models for all three sets of predictions. Figure 4 presents predicted vs. prediction error plots for urea using the augmented PLS and PACLS2/PLS models. Augmented PLS is given the spectral and concentration information from the five variable-temperature subset samples and the hybrid models are given just the five pair-wise spectral differences of the subset samples under constant- and variable-temperature conditions. Clearly the prediction abilities of the augmented hybrid models are better than that of the augmented PLS model. Since prediction errors are highly correlated

between the various methods, we use the nonparametric Wilcoxin signed-rank statistic[25] [missing reference] rather than the F-test statistic to compare prediction abilities of the various calibration methods. The improvements in the prediction errors of both hybrid models over those of the conventional PLS models are found to be significant at least at the 0.01 level in all cases. Note that the number of PLS factors used in all models changes when additional subset information is added to either PLS or the hybrid algorithms. PLS requires more factors since there are new sources of spectral variation that are present in the subset data. The hybrid algorithms require fewer PLS factors since the spectral shapes of the difference spectra are equivalent to additional CLS factors.

Although the improvements of the hybrid model predictions are significant relative to the PLS model predictions, they are still not as good as the CVSEPs of the original constant-temperature calibrations. Examination of the concentration residuals shows that the concentration residuals are correlated with temperature for both creatinine ($R^2 = 0.25$ for the PACLS2/PLS model) and NaCl ($R^2 = 0.49$). The correlation of errors with temperature inflates the prediction errors. Interestingly, the correlations of prediction errors with temperature are eliminated if we predict using models one factor less than the optimal selected during cross-validated calibration (creatinine SEP = 16 mg/dL with a 3 factor PACLS/PLS model and NaCl SEP = 20 mg/dL with a 4 factor model). Unfortunately, without prior knowledge of the concentrations in the unknowns, we would have no ability to know to reduce the number of PLS factors during prediction. Presumably, the final factor selected in the creatinine and NaCl models has resulted in an overfitting of the data by bringing in a PLS factor that is inappropriately sensitive to temperature. However, this overfitting cannot be identified from the constant-

temperature calibration data that have essentially no temperature variations in the calibration spectra.

It would be desirable to be able to update the models without requiring the performance of a full cross-validation on the augmented data. Prediction is significantly faster than the recalibration involving cross-validation for factor selection. The two sets of hybrid models allow the updating to be performed during the prediction phase of the analysis after the model has been built (see Step 2 of the hybrid prediction algorithm). The prediction results for the variable-temperature spectra using the two types of hybrids augmented in prediction are presented in Table IV. The SEP's for the CLS2/PLS hybrid augmented during prediction are indistinguishable from those of the same model augmented during cross-validated calibration. However, the SEP's from the CLS1/PLS hybrid augmented in prediction are significantly degraded from those of the same model augmented during cross-validated calibration. The degradation is a result of the PLS concentration errors in prediction not being consistent with the PLS portion of the hybrid model when one or more of the major sources of spectral variation is left out of the CLS part of the hybrid model. Thus, rapid updating in prediction is possible when the CLS portion of the hybrid model during calibration is relatively complete. When the CLS portion of the model is deficient, then in order to achieve high precision, recalibration is required.

During the augmented hybrid analysis presented above, it was not possible to predict the temperature of the variable-temperature solutions since a specific model for temperature was not available. In order for the hybrid algorithm to predict the unmodeled temperature, the spectral shape and magnitude of a pure temperature change

in the solution must be used in the augmentation portion of the PACLS/PLS algorithm. The pure shape of temperature was not present in the pair-wise subset spectral differences. However, the pure spectral shape of a temperature change in the samples can be obtained from a CLS analysis of variable-temperature spectra obtained from one of the solutions or the water solvent. A CLS analysis of the 11 variable-temperature water spectra (20 - 25° C in randomized 0.5° C increments) was performed to obtain the pure-component spectrum of the effect of a temperature change in water (see Ref. 20). Time of data collection was included along with temperature in the CLS analysis to remove linear drift effects from the temperature pure-component spectrum. In order to also model the short and long-term instrument drift, a subset of 5 spectra collected at 23° C during both constant- and variable-temperature and days were used to form pair-wise difference spectra without contamination with temperature variations. The CLS-estimated spectral shape of a temperature change and the constant-temperature subset difference spectra were added to the hybrid models in order to predict temperature changes in the variable temperature data. The resulting temperature predictions for the 26 variable-temperature samples (excluding the 5 subset sample spectra) are presented in Fig. 5, and the resulting SEP for temperature prediction was found to be 0.18° C. This result compares with the PLS cross-validated calibration for temperature prediction of 0.11° C when calibrating on the same 26 samples in the variable-temperature data set. The degradation in the temperature prediction for the hybrid versus the PLS temperature models is likely due to the very slight temperature differences that appears in the nominally constant-temperature subset spectral differences. Any presence of the shape of

a temperature change in the added subset difference spectra will reduce the net-analyte signal for temperature and, therefore, will degrade prediction precision.

CONCLUSIONS

We have introduced a new PACLS/PLS hybrid algorithm that has all the qualitative advantages of CLS, the flexibility of PLS, and the varied capabilities of the new PALCS algorithm. The hybrid also has been demonstrated to be capable of outperforming the predictive ability of PLS when predicting spectra that have new sources of unmodeled spectral variation. The improvement in prediction ability with the hybrid algorithm is highly statistically significant and is present even when the various models are based upon data from the same set of spectra. We have shown that the effects of instrument drift and solution temperature variations not in the original calibration model can be incorporated into each calibration algorithm with the use of subset samples collected during both calibration and prediction. With PLS, the subset sample spectra containing temperature and drift information during prediction are simply added back to the original calibration data and recalibration is performed to implicitly accommodate the originally unmodeled sources of spectral variation. The PACLS/PLS algorithm uses this same information in a different form. Spectral differences of pairs of the original subset sample spectra collected during calibration and prediction days are added directly to the PACLS prediction step during the hybrid prediction or during recalibration. In this manner, the PACLS portion of the hybrid algorithm is forced to explicitly ignore the sources of spectral variation present in the spectral differences. Apparently, explicitly

forcing the hybrid algorithm to ignore spectral shape information is more effective than having PLS implicitly include the spectral information during calibration.

We have found that if all the analytes are not included in the CLS calibration phase of the PACLS/PLS hybrid calibration, then better prediction ability is achieved by recalibration rather than simply updating the hybrid model during prediction. Addition of the spectral shapes during calibration of the hybrid algorithm will also improve outlier detection and will yield more realistic estimates of future prediction ability.

Since the magnitude of the added spectral shapes is not important in prediction of the original analytes and since linear combinations of the sources of spectral variation are useful for correcting the hybrid models, the required spectral shapes to add to the hybrid are generally readily obtained. However, if quantitation of an unmodeled source of spectral variation is desired, then the pure shape of the spectral interference and its quantitative magnitude must be used to augment the hybrid algorithm. We demonstrated the ability to predict unmodeled temperature of solution using the quantitative spectral shape of a temperature change of the water solvent, which was easily obtained during a separate experiment.

In order for improved predictions to be achieved, it is important that the added spectral shapes do not contain variations due to any of the analytes of interest. If analyte spectral variation is included in the augmented spectral shapes, then the net-analyte signal of the analyte will be reduced and prediction ability will be degraded. Therefore, care must be exercised in generating the augmented spectral shapes.

The capabilities of the hybrid algorithm extend beyond accommodating the presence of an unmodeled component during prediction. We can also use the hybrid

algorithm to correct for short and long-term system drift. This has been accomplished with the use of a set of subset samples in this paper, but might be more efficiently preformed with the use of a stable repeat sample measured during the calibration and prediction phases of the analysis. Mean-centered repeat spectra can then augment the calibration or prediction. Since linear combinations of the sources of spectral variation are also effective in correcting the models and since magnitudes of the vectors are not important, an eigenvector analysis of the repeat spectra may be used in the augmentation procedure. In this manner, only the most important eigenvectors required for prediction could be added to the hybrid to minimize the detrimental effect of added spectral shapes on the net-analyte signal. Another important use of the hybrid algorithm would be to transfer multivariate models between various spectrometers using subset samples measured on the primary and any secondary spectrometers. Paired spectral differences obtained from these subset spectra could then be used to augment the hybrid algorithm to correct the primary calibration model for the spectral differences of the secondary spectrometer. Future papers from our group will demonstrate these capabilities of the hybrid algorithm and will describe new methods that allow accurate updating of the multivariate models during prediction without the need to recalibrate even when only calibration information is available about a single analyte in multicomponent systems.

ACKNOWLEGEMENTS

The authors would like to acknowledge Howland D. T. Jones for making the samples and collecting the spectral data and Edward V. Thomas for providing the experimental design for the sample concentrations.

[1] W. Lindberg, J.-A. Persson, and S. Wold, Anal. Chem. **55**, 643 (1983).

[2] D. M. Haaland and E. V. Thomas, Anal. Chem. **60**, 1193 (1988).

[3] D. M. Haaland and E. V. Thomas, Anal. Chem. **60**, 1202 (1988).

[4] E. V. Thomas and D. M. Haaland, Analytical Chemistry **62**, 1091-1099 (1990).

[5] P. M. Fredericks, J. B. Lee, P. R. Osborn, and D. A. Swinkels, App. Spectrosc. **39**, 303 (1985).

[6] P. M. Fredericks, J. B. Lee, P. R. Osborn, and D. A. Swinkels, App. Spectrosc. **39**, 311 (1985).

[7] D. M. Haaland, R. G. Easterling, and D. A. Vopicka, Appl. Spectrosc. **39**, 73 (1985).

[8] Brown

[9] D. M. Haaland, "Multivariate Calibration Methods Applied to Quantitative FT-IR Analyses," in Practical Fourier Transform Infrared Spectroscopy, J. R. Ferraro and K. Krishnan, Eds., (Academic Press, New York, 1989) Chap. 8, pp. 396-468.

[10] David M. Haaland, Ling Han, and Thomas M. Niemczyk, Appl. Spectrosc. **53**, 390-395 (1999).

[11] D. W. T. Griffith, Appl. Spectrosc. **50**, 59 (1996).

[12] M. B. Esler, D. W. T. Griffith, S. R. Wilson, and I. P. Steele, Anal. Chem. **72**, 206 (2000).

[13] M. B. Esler, D. W. T. Griffith, S. R. Wilson, and I. P. Steele, Anal. Chem. **72**, 216 (2000).

[14] D. M. Haaland, W. B. Chambers, M. R. Keenan, and D. K. Melgaard, "Improved Multivariate Calibration Methods for Quantitative ICP-AES Analyses," accepted for publication in Appl. Spectrosc. **54** (2000).

[15] J. C. Ivaldi and T. W. Barnard, Spectrochim. Acta, **48B**, 1265-1273 (1993).

[16] J. A. Morales, E. H. van Veen, and M. T. C. de Loos-Vollerbregt, Spectrochim. Acta Part **B 53**, 683 (1998).

[17] H. Marten and T. Naes, "Multivariate calibration by data compression," in *Near-infrared technology in agricultural and food industries,* P. C. Williams and K. Norris, Eds., Am. Assoc. Cereal Chem., St. Paul, Minnesota, pp. 57-87 (1987).

[18] H. Marten and T. Naes, *Multivariate Calibration*, John Wiley & Sons, New York, pp. 168-213 (1989).

[19] D. M. Haaland and D. K. Melgaard, "New Prediction-Augmented Classical Least Squares (PALCS) Methods: Application to unmodeled Interferents," accepted for publication in Appl. Spectrosc. **54** (2000).

[20] D. M. Haaland, Appl. Spectrosc. **54**, 246 (2000).

[21] D. M. Haaland and H. D. T. Jones, "Multivariate Calibration Applied to Near-Infrared Spectroscopy for the Quantitative Analysis of Dilute Aqueous Solutions," 9th International Conference on Fourier Transform Spectroscopy, J. E. Bertie and H. Wisser, Editors, Proc. SPIE Vol. **2089**, 448 (1993).

[22] E. V. Thomas and N. Ge, Technometrics **42**, 168 (2000).

[23] C. L. Lawson and R. J. Hanson, "*Solving Least Squares Problems*," Prentice-Hall, Englewood Cliffs, NJ (1974).

[24] H. Mark, Anal. Chem. **59**, 790 (1987).

[25] Wilcoxin signed-rank statistic

Table I. Cross-validated calibration results for constant-temperature aqueous solutions.

Component	PLS			CLS2/PLS			CLS1/PLS		
	CVSEP	R^2	Factors	CVSEP	R^2	Factors	CVSEP	R^2	Factors
Urea (mg/dL)	16	0.9997	11	15	0.9998	5	15	0.9998	9
Creatinine (mg/dL)	15	0.9997	12	14	0.9997	6	14	0.9998	10
NaCl (mg/dL)	15	0.9997	13	18	0.9996	6	18	0.9996	10

Table II. True prediction results of variable-temperature data using various constant-temperature models.

Component	PLS			CLS2/PLS			CLS1/PLS		
	SEP	R^2	Factors	SEP	R^2	Factors	SEP	R^2	Factors
Urea (mg/dL)	250	0.9264	11	322	0.8776	5	331	0.8705	9
Creatinine (mg/dL)	238	0.9129	12	289	0.8724	6	265	0.8926	10
NaCl (mg/dL)	225	0.9333	13	266	0.9066	6	288	0.8910	10

Table III. True prediction results of variable-temperature data using constant-temperature models with subset spectral information added during calibration.

Component	PLS			PACLS2/PLS			PACLS1/PLS		
	SEP	R^2	Factors	SEP	R^2	Factors	SEP	R^2	Factors
Urea (mg/dL)	43	0.9978	11	17	0.9997	4	17	0.9997	8
Creatinine (mg/dL)	36	0.9980	15	19	0.9995	4	23	0.9992	8
NaCl (mg/dL)	39	0.9980	15	27	0.9990	5	27	0.9990	9

Table IV. True prediction results of variable-temperature data using constant-temperature models with subset spectral information added during prediction.

Component	PACLS2/PLS			PACLS1/PLS*		
	SEP	R^2	Factors	SEP	R^2	Factors
Urea (mg/dL)	22	0.9994	4	100	0.9883	9
Creatinine (mg/dL)	18	0.9995	4	44.9	0.9969	10
NaCl (mg/dL)	26	0.9991	5	479	0.915	10

* Mean-centered model

FIGURE CAPTIONS

Figure 1. A) Mean spectrum of the constant-temperature data, B) mean-centered constant-temperature spectra, and C) mean-centered variable-temperature spectra.

Figure 2. Cross-validated calibration results for creatinine using the new CLS2/PLS calibration model that includes the three analytes, water, and time of data collection in the CLS portion of the hybrid algorithm. The solid line represents the line of identity.

Figure 3. Creatinine predictions of using the constant-temperature CLS2/PLS model applied to the variable-temperature data. The solid line represents the line of identity.

Figure 4. Urea prediction errors for the variable-temperature data using the constant-temperature PLS model augmented with the 5 subset variable temperature data (open triangles) and the PACLS2/PLS model augmented during calibration with the 5 paired subset sample difference spectra from the same 5 subset samples (X's).

Figure 5. Temperature predictions of the variable-temperature data using the constant-temperature CLS2/PLS model augmented during calibration with the 5 paired 23° C subset sample difference spectra and the pure spectrum of a temperature change in water obtained from variable temperature pure water spectra. The solid line represents the line of identity.

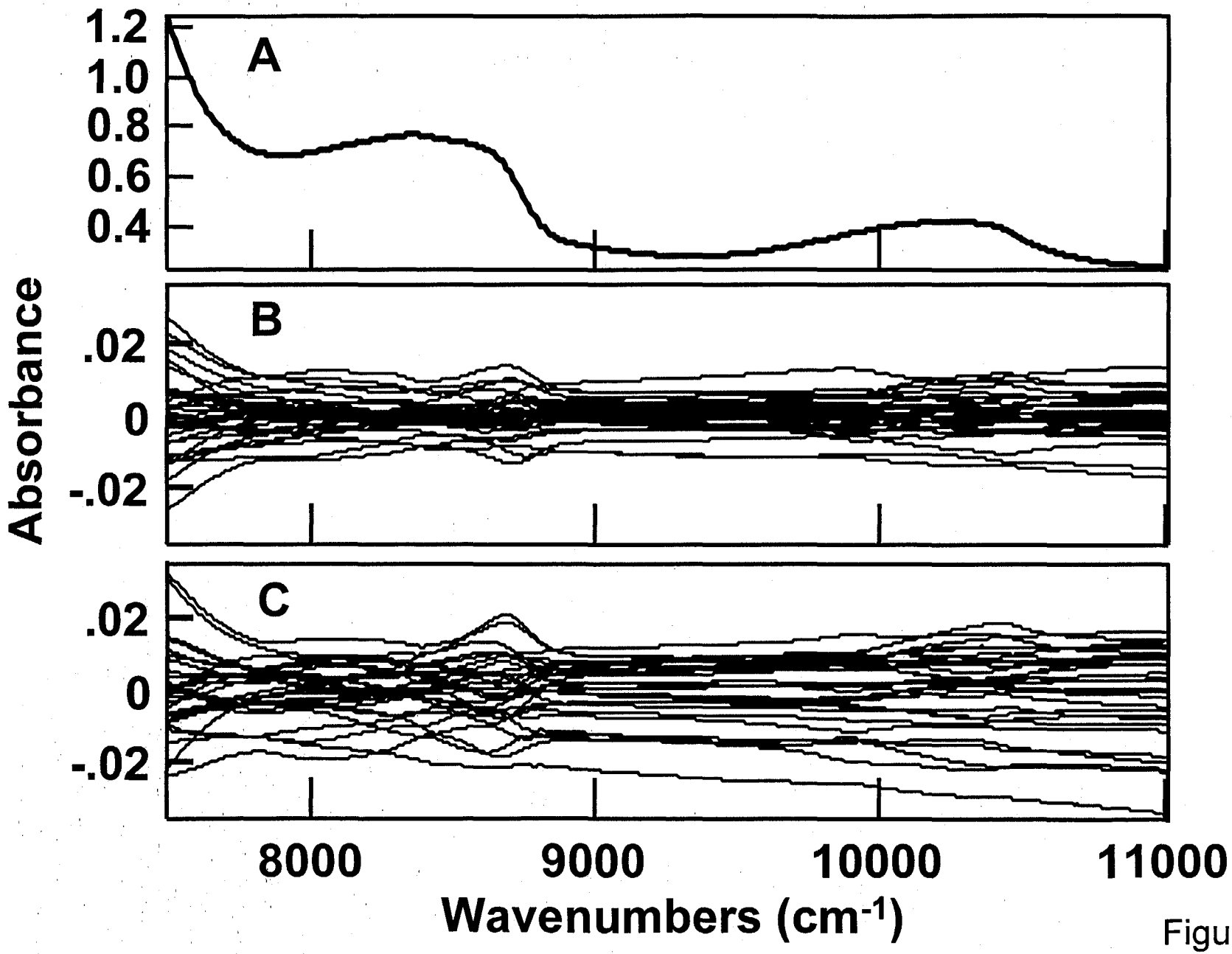
1.2
1.0
0.8
0.6
0.4
A
B
.02
0
-.02
C
.02
0
-.02
Absorbance
8000
9000
10000
11000
Wavenumbers (cm-1)

Figure 1

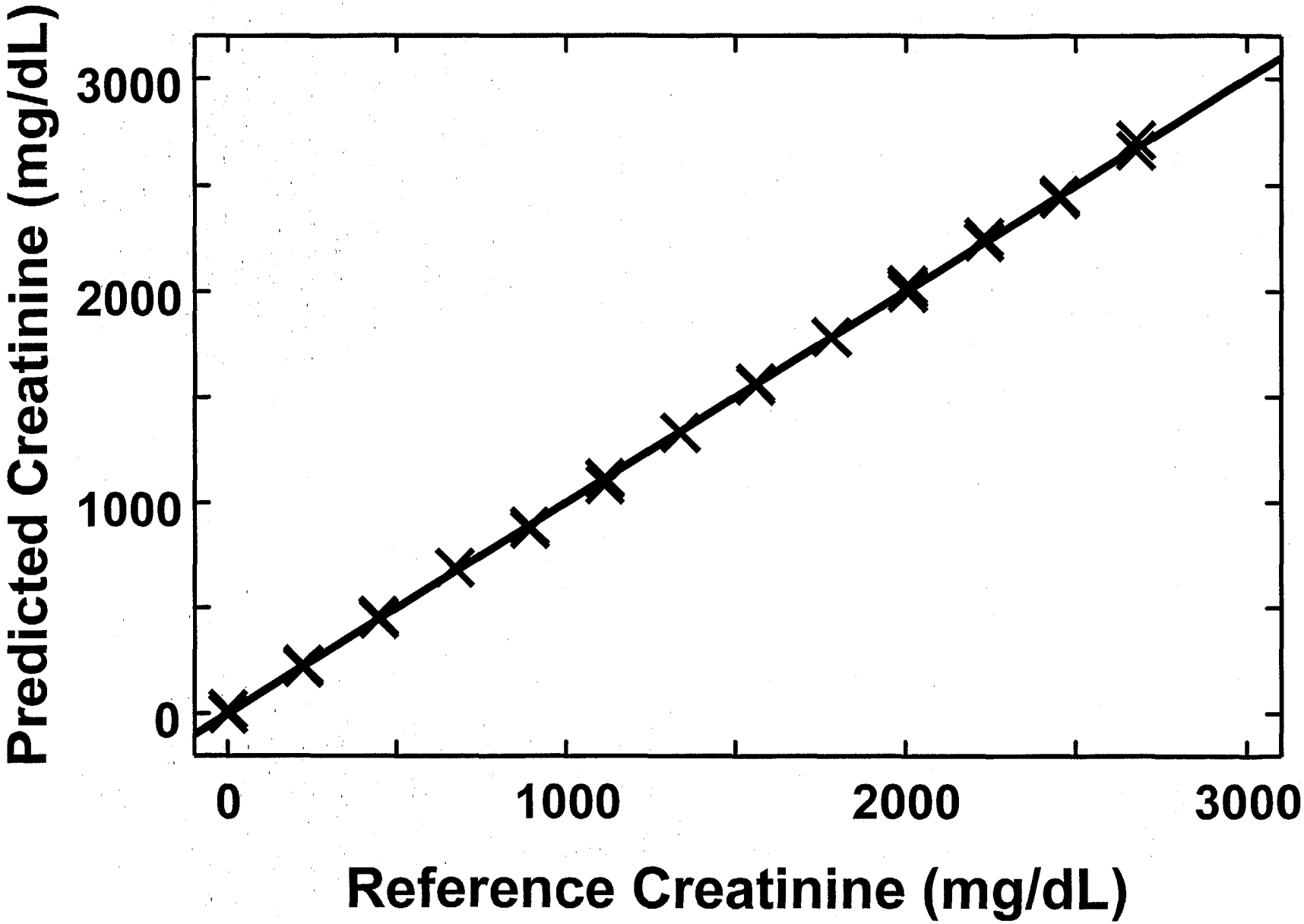
Predicted Creatinine (mg/dL)
3000
2000
1000
0
0
1000
2000
3000
Reference Creatinine (mg/dL)

Figure 2

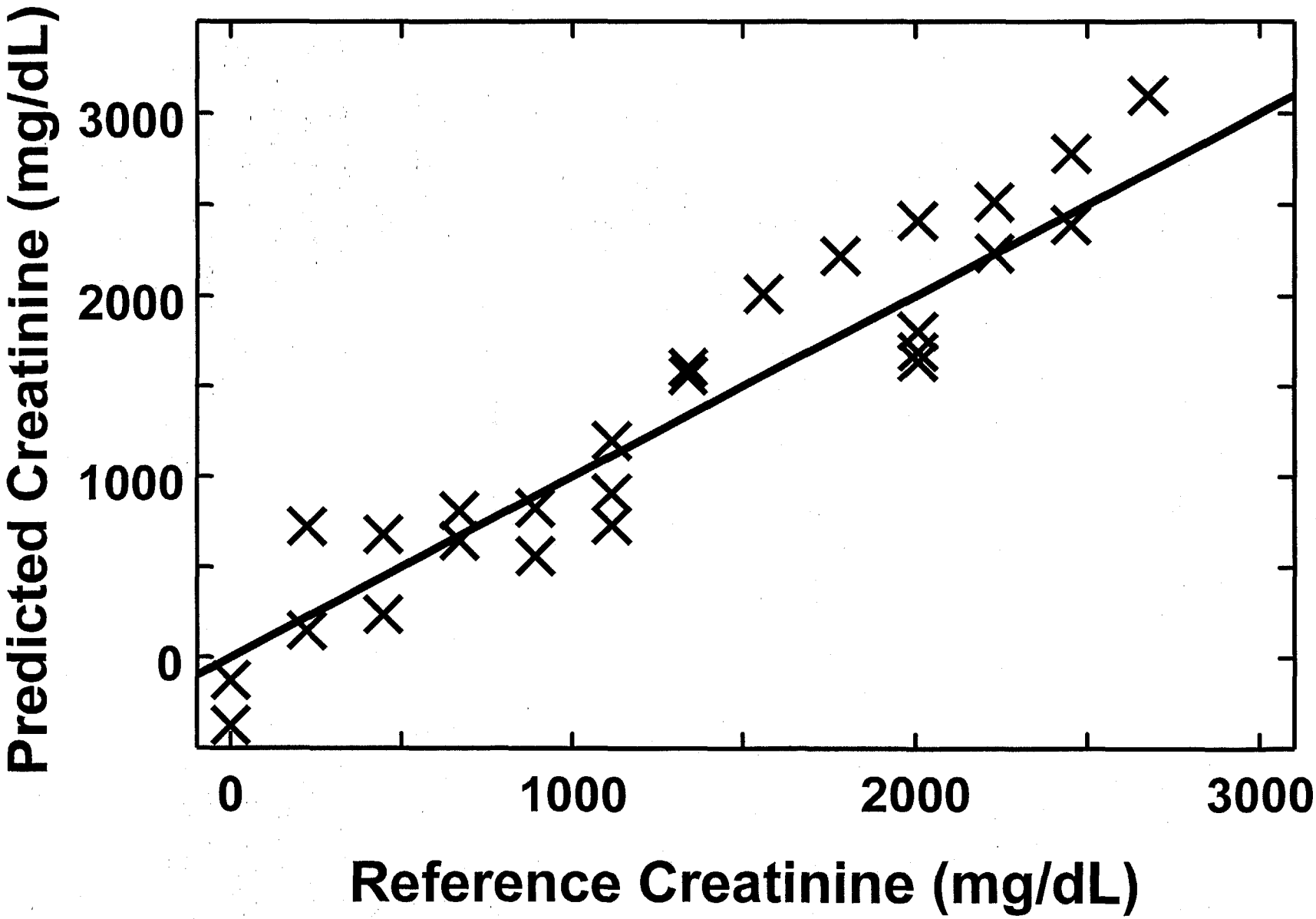
Predicted Creatinine (mg/dL)
3000
2000
1000
0
0
1000
2000
3000
Reference Creatinine (mg/dL)

Figure 3

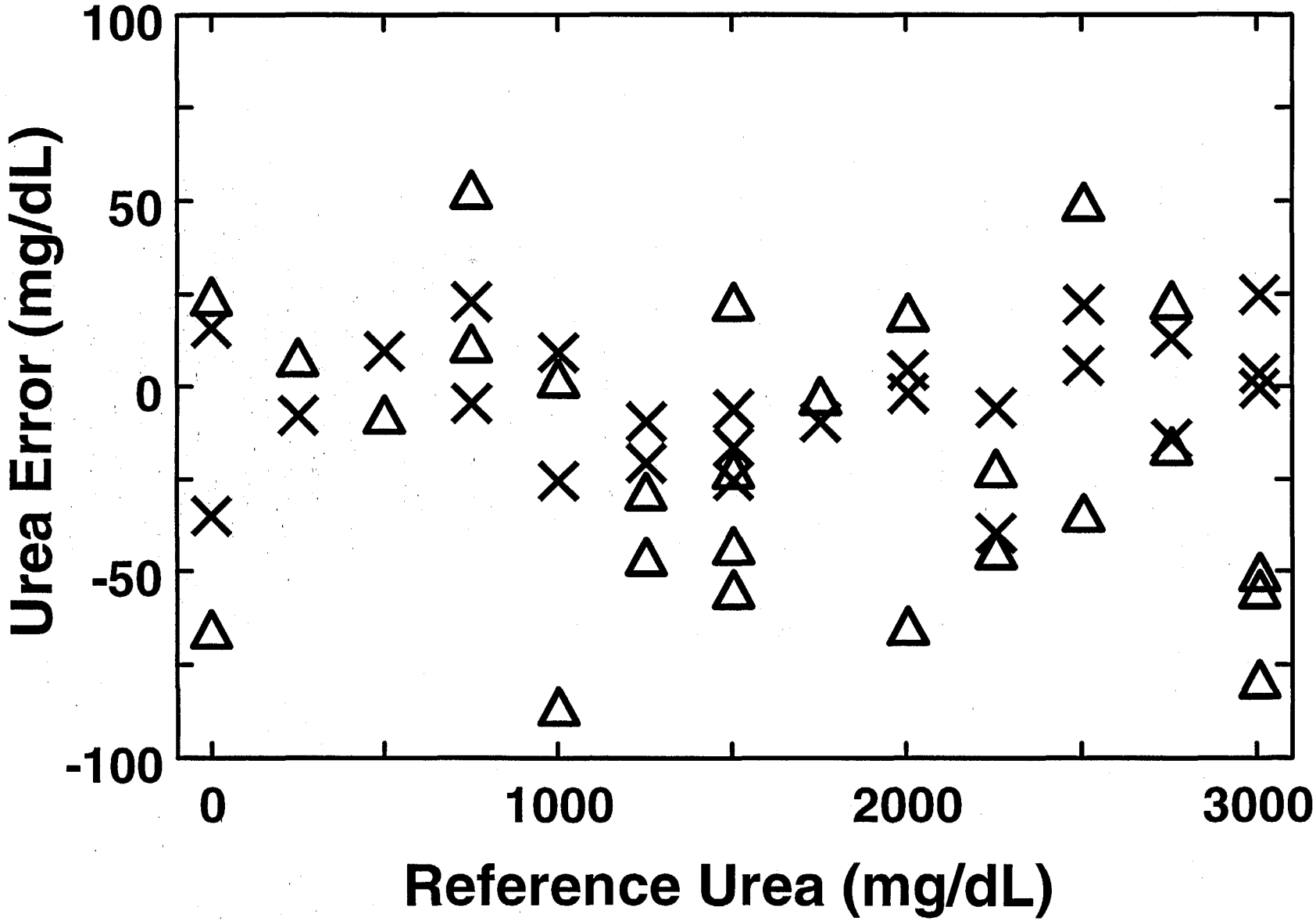
100
50
0
-50
-100
Urea Error (mg/dL)
0
1000
2000
3000
Reference Urea (mg/dL)

Figure 4

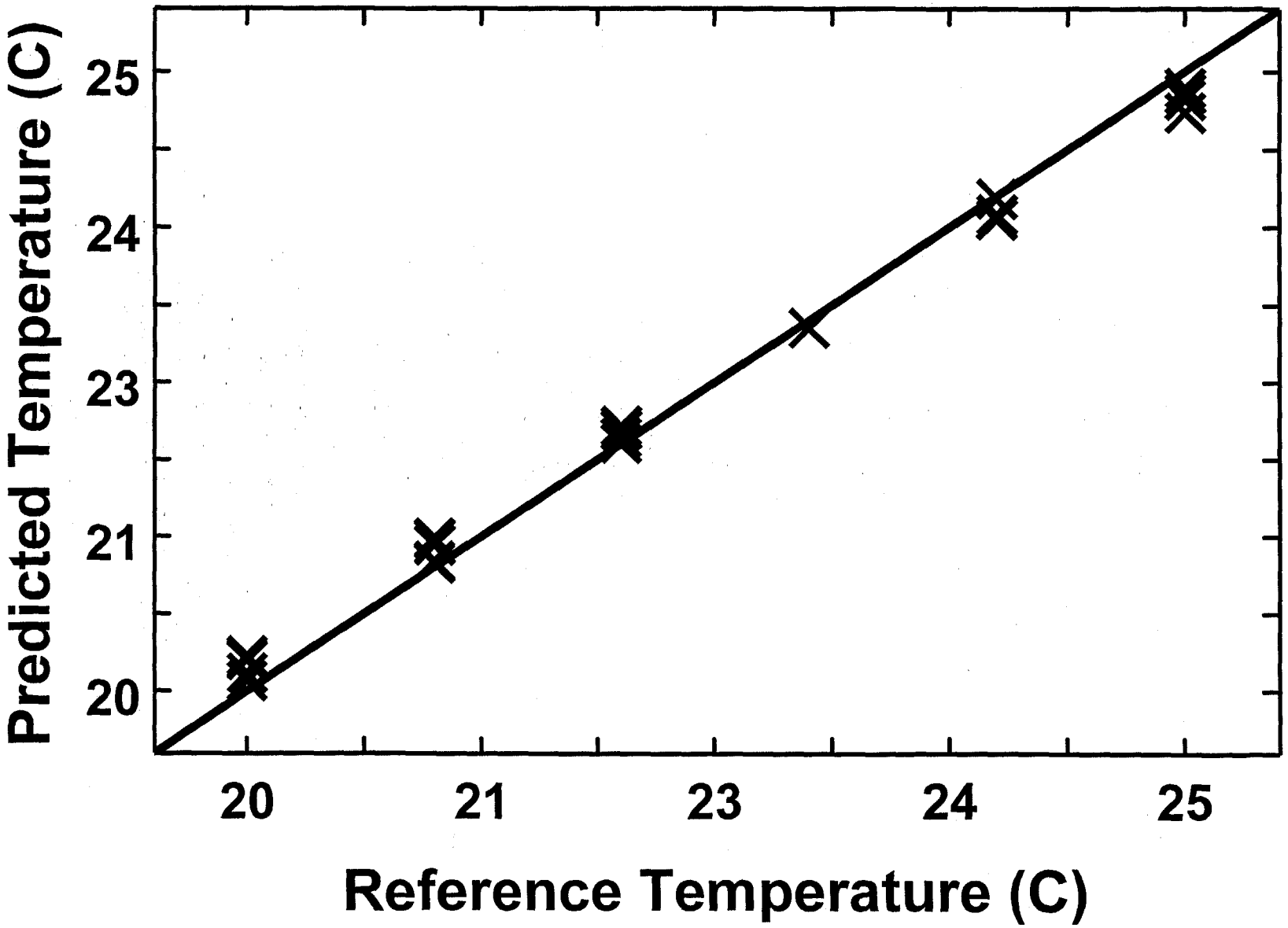
Predicted Temperature (C)
25
24
23
21
20
20
21
23
24
25
Reference Temperature (C)

Figure 5