# CVD Experiment Designer & Analyzer

**v2.1, June 2025**

This document provides an overview of the cvd\_DoE\_v2.html web-based tool. The tool is designed to assist researchers in designing, analyzing, and optimizing experiments based on Response Surface Methodology (RSM) for Chemical Vapor Deposition (CVD) processes.

### 1. Overview & Workflow

The application guides the user through a structured workflow for performing a Design of Experiments (DoE) analysis for a CVD process. The primary goal is to find the optimal set of process parameters (factors) to maximize one or more desired outcomes (responses), such as film thickness and process yield.

The user workflow is as follows:

1. **Design Experiment**: The user defines the process factors, their minimum and maximum operating values, and selects a DoE method. The tool provides default factors for Precursor Concentration, Furnace Temperature, N2/O2 Flow Ratio, Pressure, and Deposition Time.
2. **Generate Plan**: The tool generates a detailed experimental plan, listing the specific combination of factor settings for each run.
3. **Enter Data**: The user performs the experiments and enters the measured responses (Thickness and Yield) into the data table. A feature to fill the table with simulated data based on a pre-defined mathematical model is also available for validation and demonstration purposes.
4. **Analyze & Optimize**: The tool automatically performs a regression analysis to create a mathematical model linking the factors to the responses. It calculates key statistical metrics, predicts optimal conditions, and visualizes the relationships using contour plots.
5. **Export**: A summary report of the data, models, and predicted optima can be exported as a PDF document.

### 2. Design of Experiments (DoE) Methods

The tool implements three standard Response Surface Methodology (RSM) designs. RSM is a collection of statistical and mathematical techniques useful for developing, improving, and optimizing processes.

#### 2.1. Full Factorial Design

A full factorial experiment measures the response at all possible combinations of the factor levels. For *k* factors, each at 2 levels (min and max), this design consists of runs.

* **Purpose**: Excellent for identifying the main effects of factors and their interactions.
* **Implementation**: The generateFactorial function creates this plan. It is primarily suited for fitting first-order models with interaction terms and serves as the foundation for the Central Composite Design.

#### 2.2. Box-Behnken Design (BBD)

A Box-Behnken design is an efficient, three-level design used for fitting second-order (quadratic) models. It does not contain runs at the extreme vertices of the experimental space, which can be advantageous when these points represent expensive or unsafe operating conditions.

* **Purpose**: To fit a quadratic model and find optima without needing to test extreme factor combinations.
* **Structure**: The design is constructed by combining two-level factorial designs with incomplete block designs. The generateBoxBehnken function implements this by creating runs where factors are at their center points while pairs of other factors are at their min/max levels.
* **Requirement**: Requires at least 3 varying factors.

#### 2.3. Central Composite Design (CCD)

A CCD is the most common design used for building second-order models. It is highly efficient and flexible.

* **Purpose**: To efficiently estimate a second-order model, allowing for the detection of curvature in the response surface and the location of optimal process settings.
* **Structure**: The generateCentralComposite function builds the design from three types of points:
  1. **Factorial Points**: The corners of the experimental space (coded as -1 and +1).
  2. **Center Points**: Replicate runs at the center of the domain (coded as 0).
  3. **Axial (or Star) Points**: Points along the axes of the factors, set at a distance ‘α’ from the center, which enable the estimation of quadratic terms.
* **CCD Type (α value)**:
  + **Face-Centered (α=1.0)**: The axial points are located on the “faces” of the factorial cube, confining all runs within the specified min/max range.
  + **Rotatable (α = k¹/⁴)**: Here, α is calculated from the number of factors (k) to ensure the prediction variance is constant at points equidistant from the center. This can result in axial points outside the min/max range.

### 3. Mathematical Foundation

#### 3.1. Coding of Variables

To simplify calculations, real-world factor values (e.g., 500°C to 900°C) are “coded” into a dimensionless range from -1 to +1.

The uncodeValue function reverses this process.

#### 3.2. The Model

The application uses a second-order polynomial model. For two factors, and , the model is:

Where is the predicted response (e.g., Thickness), is the intercept, are linear coefficients, are interaction coefficients, are quadratic coefficients, and is the error.

#### 3.3. Model Fitting: Method of Least Squares

The coefficients () are determined using the method of least squares, solved via the matrix equation:

Where **b** is the vector of coefficients, **X** is the design matrix, and **y** is the vector of observed responses.

### 4. Analysis of Statistical Parameters

The buildModel function calculates several key statistics to evaluate the quality, significance, and predictive capability of the fitted second-order polynomial model.

#### 4.1. Interpretation of Model Coefficients

The coefficients in the regression model quantify the effect of each factor on the response. Because they are calculated using coded values, their magnitudes can be directly compared.

* **Intercept ()**: The predicted response when all factors are at their center point (coded value 0).
* **Linear Term (e.g., )**: Represents the main effect of a factor. A positive coefficient means the response increases as the factor increases. A negative coefficient means the response decreases.
* **Interaction Term (e.g., )**: Indicates that the effect of one factor depends on the level of another.
  + A **positive** interaction (Temp \* Pressure) means the factors have a synergistic effect; the effect of temperature on the response is stronger at higher pressures.
  + A **negative** interaction means the factors have an antagonistic effect; the effect of one is dampened by the other.
* **Quadratic Term (e.g., )**: Represents the curvature in the response surface for a single factor. This is crucial for finding an optimum.
  + A **negative** quadratic term (e.g., Conc \* Conc or ) is highly desirable when maximizing a response. It indicates that the response increases with the factor up to a certain point, and then begins to decrease, creating a “hump” or maximum on the response surface. The model can locate the peak of this hump.
  + A **positive** quadratic term indicates the opposite: the response decreases and then increases, forming a “trough” or a minimum. This would be desirable if the goal were to minimize the response.

#### 4.2. R-Squared ()

* **Definition**: Also known as the coefficient of determination, R-Squared measures the proportion of the total variation in the observed response variable that is explained by the model.
* **Formula Used**: , where is the sum of squared residuals and is the total sum of squares.
* **Interpretation**: The value ranges from 0 to 1 (or 0% to 100%). An of 0.90 means that 90% of the response’s variability is accounted for by the model’s factors. A higher is generally better.

#### 4.3. Adjusted R-Squared (Adj )

* **Definition**: A modified version of R-Squared that is adjusted for the number of predictors (terms) in the model.
* **Interpretation**: The Adjusted increases only if the new term improves the model more than would be expected by chance. It is always lower than and is considered a more reliable measure of model fit as it penalizes “overfitting.”

#### 4.4. Predicted R-Squared (Pred )

* **Definition**: This statistic measures how well the model is expected to predict responses for new, unseen data points.
* **Interpretation**: A high Predicted indicates good predictive power. It should be in “reasonable agreement” with the Adjusted . A large difference (e.g., > 0.2) between them suggests the model may be overfit. This is one of the most important metrics for validating a model’s practical utility.

#### 4.5. PRESS (Predicted Residual Sum of Squares)

* **Definition**: A form of cross-validation used to calculate Predicted . It is calculated by systematically removing each data point, fitting the model without it, and summing the squared errors of predicting the left-out point.
* **Interpretation**: A lower PRESS value is better, indicating smaller prediction errors. It is the foundation for calculating Pred .

#### 4.6. Adequate Precision (Adeq. Precision)

* **Definition**: This statistic measures the signal-to-noise ratio. It compares the range of the predicted values at the design points to the average prediction error.
* **Interpretation**: It indicates whether the model provides sufficient signal to be used for navigating the design space. **A ratio greater than 4 is essential and considered a minimum for a valid model.** A value below 4 suggests the model is just fitting noise and should not be used for prediction or optimization.

#### 4.7. Standard Deviation (Std. Dev.) & C.V. %

* **Definition**: The Standard Deviation is the square root of the residual mean square (error variance) and represents the typical error in the units of the response. The Coefficient of Variation (C.V. %) expresses this error as a percentage of the response mean.
* **Interpretation**: Lower values are better. A low C.V. % (e.g., < 10-15%) generally indicates good model reproducibility.

#### 4.8. Coefficient Statistics (Std. Error & 95% CI)

* **Definition**: These statistics assess the significance of each individual term in the model. The 95% Confidence Interval (CI) gives the range where the true coefficient value likely lies.
* **Interpretation**: If the 95% CI for a term’s coefficient **contains zero**, that term is generally not statistically significant at the 5% level. Its effect cannot be distinguished from random experimental noise.

### 5. Optimization

#### 5.1. Single-Response Optimization

The findGlobalOptimum function seeks the combination of factor settings that maximizes a single response (Thickness or Yield) using a random walk numerical optimization algorithm.

#### 5.2. Multi-Response Optimization (Weighted Score)

The tool optimizes multiple responses using a weighted sum approach. The user assigns an importance weight to each response using sliders. The findCombinedOptimum function then maximizes a combined score to find a “compromise” setting that provides the best overall outcome.

Where is the weight for each response and is the normalized predicted value for that response.

### 6. Visualization

The tool generates 2D contour plots to help visualize the response surface. A star symbol (★) on the plot marks the location of the predicted mathematical optimum for that response. When creating a 2D plot for a design with more than two factors, the other factors are held constant at their center point (average) value.

### 7. About the File

* **Filename**: cvd\_DoE\_v2.html
* **Version**: 2.1
* **Date**: June 28, 2025